MAPPING FUTURE LAND COVER CHANGE OVER LARGE AREAS OF THE UNITED STATES USING DECISION TREES

A Thesis
Presented to
The Academic Faculty

by

Felipe Sant'Anna Dias

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Environmental Engineering in the
School of Civil and Environmental Engineering

Georgia Institute of Technology
May 2015

MAPPING FUTURE LAND COVER CHANGE OVER LARGE AREAS OF THE UNITED STATES USING DECISION TREES

Approved by:

Dr. Marc Stieglitz, Committee Chair
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Greg Turk
School of Interactive Computing
Georgia Institute of Technology

Dr. Mustafa Aral
School of Civil and Environmental Engineering
Georgia Institute of Technology

Date Approved: 21 april 2015

To my wife,

Renee M. Dias,

for all her love and support.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Marc Stieglitz, for all his commitment and guidance. He has dedicated a tremendous amount of time and effort to mentoring this thesis work. I would like to thank the rest of my committee members, Dr. Greg Turk and Dr. Mustafa Aral, for their time and consideration.

I also wish to thank my lab and research coworker, Yashika Agarwalla, for all the help and assistance provided to me during my research period.

I would like to thank my wife, Renee Dias, for providing unconditional support and love during this stressful time. I know that without her I would not be able to achieve my goals.

Lastly, I want to thank my parents, Fernando and Izaline Dias, for supporting my studies and providing me with the opportunity to attend Graduate School at the Georgia Institute of Technology.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Climate is one of the primary factors that control vegetation distribution and therefore it is expected that the effects of climate change will have a significant impact on the natural land cover. Numerous models, like Dynamic Global Vegetation Models (DGVMs), have been developed to project the potential shift in vegetation distribution under rapid climate change. However, those models present a great constraint on the amount of data that can be processed, making it unable to simulate vegetation distribution over large areas with an exceedingly high resolution. To overcome this limitation, new alternative methods have been proposed to study vegetation distribution and natural land cover classification using statistical techniques.

Machine Learning is a scientific discipline that utilizes computer algorithms to learn patterns and statistical rules, based on present correlation defined by a training set, that can be applied to predict new information. Among different machine learning algorithms, the Decision Tree model has been widely used to classify present land cover and account land use modifications, making it a suitable model to statistically learn present vegetation distribution pattern, in order to be applied to predict future shifts in the biogeography with climate change.

The decision tree algorithm applied in this work is the C5.0 classification tree, which provides classified images of future vegetation cover at four large sites in the US; a region including the Jemez and Santa Fe mountains located at north central New Mexico, a region of the Blue Mountains in Oregon, a region of the North Cascades located at the northwest Washington, and totality of the Wyoming State. The training data used to generate current vegetation cover include 2001 USGS Land

Cover maps, 50 years of mean annual temperature and annual precipitation for the period 1950-2000, and Digital Elevation Model together with aspect and slope data. Future climate data was generated using Model E2 version of the Goddard Institute for Space Studies (GISS) General Circulation Model (GCMs) downscaled and bias corrected for the current climate data. Four future climate scenarios, RCP 2.6, RCP 4.5, RCP 6.0 and RCP 8.5, were used for generating the future climate for the target year of 2070 (average for 2061-2080 period).

The model performed well for all four locations, achieving prediction accuracies for current land cover of 83%, 85%, 82% and 80% respectively for New Mexico, Oregon, Washington and Wyoming sites. Each site presented different types of modifications for future terrestrial ecosystems.

CHAPTER I

INTRODUCTION

Climate is one of the primary factors that controls vegetation and plant species distribution, therefore it is expected that the effects of climate change will have a significant impact, either negative or positive, on the natural land cover (Cramer et al., 2001; Pearson and Dawson, 2003; Huntley et al., 2004). Besides affecting the species distribution, climate change can also modify the biogeochemical cycles, change community structure through changes in species composition and abundance (Mouillot et al., 2002) and affect the ecosystem net primary production (Melillo et al., 1993). On top of the fact that climate change impacts the natural vegetation distribution, changes in land cover and land use provoked by human action can also aggravate even further the change in climate by altering biogeophysical characteristics, such as the surface albedo and potential evapotranspiration, and thus altering surface-atmosphere energy exchanges (Brovkin et al., 2004; Petit et al., 2001; Lambin et al., 2003; Pielke, 2005; Feddema et al., 2005; Bonan, 2008).

The effect of climate change over the biogeography have motivated many researchers to develop models that relate species distributions and climate and use them to simulate how these distributions may be altered in response to potential future climate scenarios (Huntley et al., 2004; Mouillot et al., 2002; Pearson and Dawson, 2003; Cramer et al., 2001; Sala et al., 2000). Several types of models, like the Dynamic Global Vegetation Models (DGVMs), Climatic Envelope Models (CEMs) and Climate Response Surface Models, have been developed to project the potential shift in vegetation distribution under rapid climate change (Cramer et al., 2001; Huntley et al., 1995; Hijmans and Graham, 2006). From all of the different model types, the

DVGM is probably the most popular and most used among the scientific community (Cramer et al., 2001). DVGM is a computer program that simulates shifts in potential vegetation and its associated biogeochemical and biophysical cycles as a response to alterations in climate; it uses time series of climate data and ancillary data, like topography and soil characteristics, to simulate dynamics of ecosystem processes (Hickler et al., 2012). These models are commonly applied over extremely large areas in order to study the vegetation dynamics in a regional or global scale, however the resolution is often very coarse; ranging from 30 to 300 km (0.25 to 2.5 °) (Anav et al., 2010; Sitch et al., 2008; Beer et al., 2007; Krinner et al., 2005; Hickler et al., 2012; Cramer et al., 2001; Hijmans and Graham, 2006; Huntley et al., 1995). Vegetation models can also be applied in a higher resolution manner, but they are limited to small areas (Bachelet, 2001). Therefore, models like DVGMs present a great constraint on the amount of data that can be processed, making it unable to simulate vegetation distribution over large areas with an exceedingly high resolution.

To overcome the limitations and complexities of computationally expensive simulation models, new alternative methods have been proposed to study vegetation distribution and natural land cover classification using statistical techniques; based on the premise that the geographical distribution of species are statistically related to present environment and climate conditions (Austin, 2002; Guisan and Zimmermann, 2000). The Machine Learning science field presents several algorithms that learns patterns and statistical rules, based on present correlation defined by a training set, which can be applied to predict new information (Mitchell, 1997); therefore, its application would be well suitable for predicting vegetation distribution and classifying land cover (DeFries and Chan, 2000; Srivastava et al., 2012). Among different machine learning algorithms, the Decision Tree model has been widely used to classify present land cover and account land use modifications (DeFries and Chan, 2000;

Klein et al., 2012; de Colstoun et al., 2003; McIver and Friedl, 2002; Homer et al., 2004), making it a suitable model to statistically learn present vegetation distribution pattern, in order to be applied to predict future shifts in the biogeography with climate change.

Machine Learning (ML) is a scientific discipline in the field of computer science and statistics that is directly related to the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed. It encompasses automatic computing procedures based on logical or binary operations, which learn a task from a series of examples or specific training samples and uses it to make further decisions or predictions (Bishop et al., 2006). Machine learning has the capability to discover previously unknown regularities and trends in databases and also helps people to explicate, codify and reproduce their knowledge and expertise (Witten et al., 1993). There are three major research niches in machine learning: data mining, which consist of the use of historical data to extract relevant information in order to improve future decision making, task-oriented studies, which is the development and analysis of self-learning systems that improve performance in a predetermined set of tasks, for example, autonomous driving and speech recognition, and cognitive simulations, which represents the investigation and computer simulation of man learning processes related to intelligence and behavior, an example is programs that learn the users interest (Carbonell et al., 1983; Mitchell, 1997).

This thesis will study the implementation of a decision tree method, specifically the C5.0 algorithm, to provide classified images of future vegetation cover. A decision tree is a schematic tree-shaped diagram used as a decision support tool. It describes graphically the decisions to be made, the events that may occur, and the outcomes associated with combinations of those decisions and events (Witten and Frank, 2005). Decision trees are useful tools for many data mining problems, where both predictive

accuracy and the ability to analyze the model are important. They are developed using different measurements that recursively split data sets into increasingly homogeneous subsets representing class membership (Quinlan, 1986). An important advantage of decision trees is that they are structurally explicit, allowing for clear interpretation of the links between predictors and the classes (Quinlan, 2014). Also, a decision tree classification scheme offers an efficient and reliable method to classify large quantities of information in a short period of time (Witten and Frank, 2005).

Data obtained from four large sites in the United States of America will be used to train the decision tree model and then, using future climate data, testing its capability of predicting future land cover. The sites chosen are: a region including the Jemez and Santa Fe mountains located in northern central New Mexico, a region of the Blue Mountains in Oregon, a region of the North Cascades located in northwest Washington, and the totality of Wyoming State. All sites present an extremely high resolution, ranging from 150 to 600 meters, which is much finer when compared to the resolution used in the traditional vegetation models described earlier, like the DVGM.

The data used to train the decision tree for current vegetation cover include the 2001 USGS Land Cover maps, 50 years of mean annual temperature and annual precipitation for the period 1950—2000, and Digital Elevation Model together with aspect and slope data. Future climate data generated using Model E2 version of the Goddard Institute for Space Studies (GISS) General Circulation Model (GCMs) and downscaled and bias corrected to the current climate, was used to predict new classified land cover maps.

Four different future climate change scenarios were used for generating future climate surface data for the target year of 2070 (average for the 2061-2080 period).

The climate change scenarios, referred to as Representative Concentration Pathways (RCPs), are climate simulations for the twenty-first century (2000—2100) carried out under the framework of the Coupled Model Intercomparison Project Phase 5 (CMIP5) of the World Climate Research Program (IPCC, 2013). The RCPs are consistent sets of projections of only the components of radiative forcing meant to serve as input for climate and atmospheric chemistry modeling and pattern scaling as part of the preparatory phase for the development of new scenarios for the IPCC's Fifth Assessment Report and beyond (IPCC, 2013).

The RCPs are four independent pathways developed by four individual modeling groups. They are labeled according to the approximate target radiative, forcing at year 2100 relative to pre-industrial climate conditions (Meinshausen et al., 2011). Radiative forcing, expressed as Watts per square meter, is the additional energy taken up by the Earth's system due to the enhanced greenhouse effect; it can be defined as the difference in the balance of energy that enters the atmosphere and the amount that is returned to space compared to the pre-industrial conditions (Ramaswamy et al., 2001). The net radiative forcing is determined by both positive forcing from greenhouse gases and negative forcing from aerosols, though the dominant factor across the scenarios is the forcing from CO2 concentration. The four RCPs, with forcing values from 2.6 to 8.5 W/m, are: the lowest forcing level scenario RCP 2.6 (Van Vuuren et al., 2011), two median range scenarios RCP 4.5 (Thomson et al., 2011) and RCP 6.0 (Masui et al., 2011), and the business-as-usual scenario RCP 8.5 (Riahi et al., 2011). Radiative forcing agents present include: time-varying well-mixed greenhouse gases emissions (CO2, CH4 and N2O), ozone, tropospheric aerosols (sulfates, nitrates, black carbon and organic carbon), stratospheric water vapor from methane oxidation, a parameterized indirect effect of aerosols on clouds, soot effect on snow and ice albedos, anthropogenic land use changes, volcanic aerosols, solar

irradiance, and Earth orbital parameters (Schmidt et al., 2014).

The figure 1 and figure 2 are, respectively, projections for the CO2 equivalent concentration and for the radiative forcing levels of each of the four RCPs.

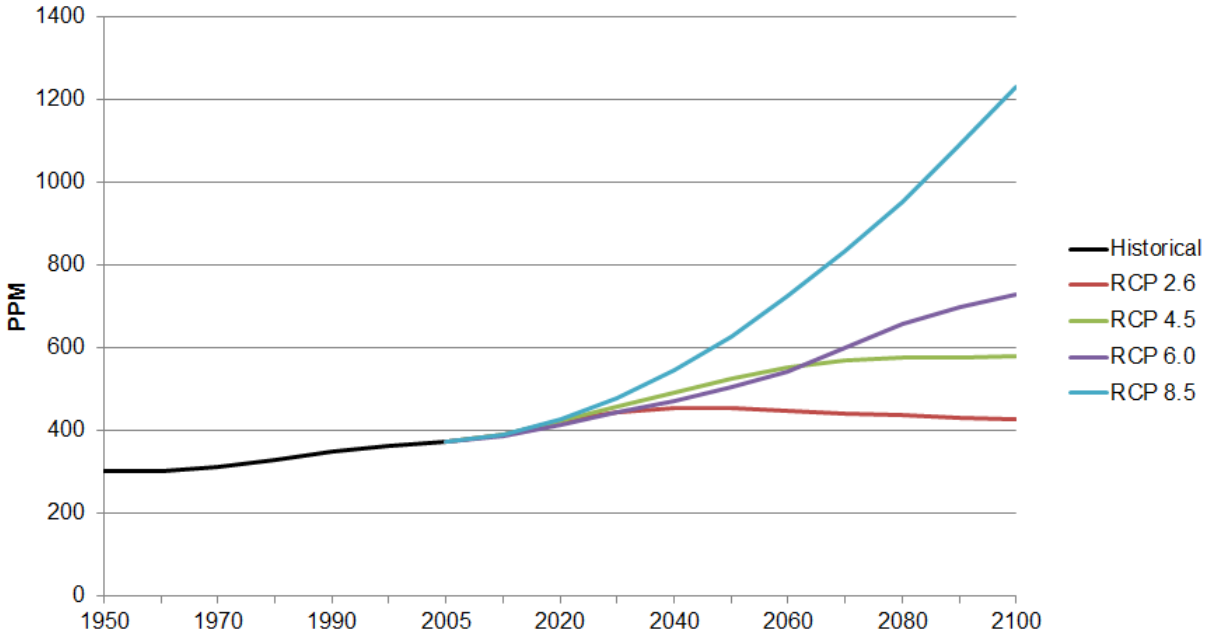

Figure 1: RCPs Climate Scenarios - Atmospheric CO2-equivalent Concentration

The C5.0 algorithm is explained in more detail in chapter II. Descriptions of the four studied sites are presented in chapter III. Data and Software utilized in this research are detailed in chapter IV. The methodology and the results are presented in chapter V and VI, respectively. Finally, Chapter VII gives a comparison of the two ML techniques followed by discussions and conclusions.

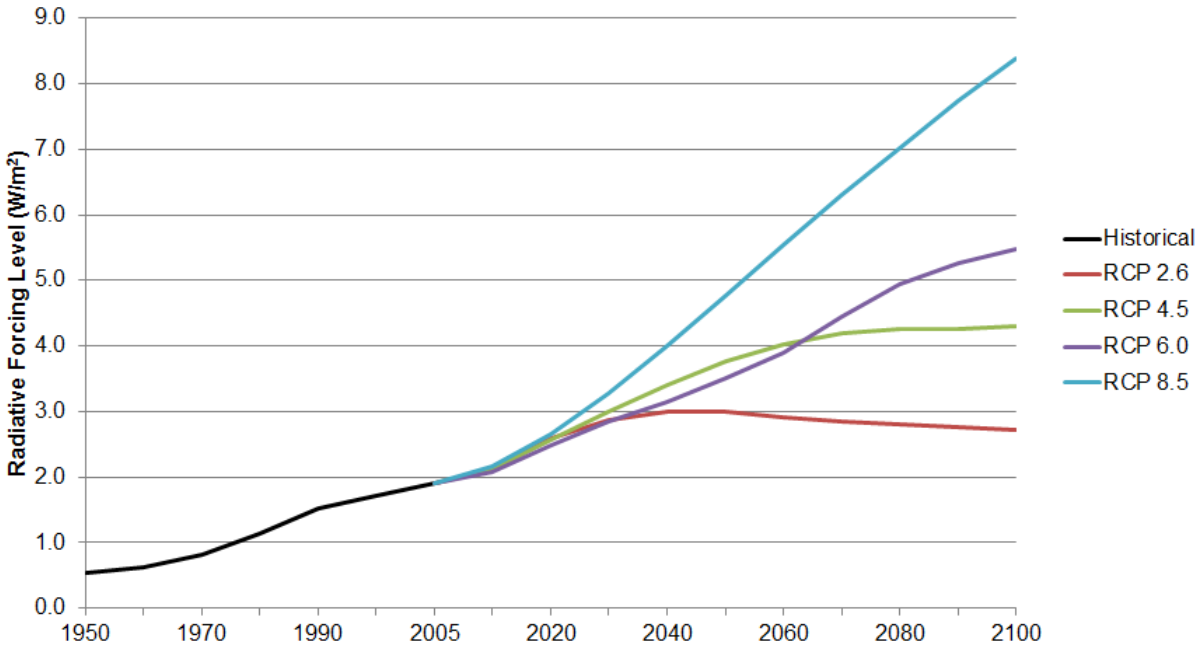

Figure 2: RCPs Climate Scenarios - Radiative Forcing Levels

CHAPTER II

C5.0 DECISION TREE ALGORITHM

2.1 Introduction and Structure

C5.0 is a machine learning algorithm that generates decision trees from a dataset; it is widely used to solve data mining tasks. The algorithm was developed by Ross Quinlan as an extension of his previous decision tree model, the C4.5. It carries several improvements from its predecessor, such as; a faster, highly optimized and more memory efficient algorithm; the boosting option, a technique for constructing multiple tree classifiers to improve model accuracy; the winnowing option, that discards less relevant attributes, which is useful for datasets with a higher amount of attributes (RuleQuest, 2009).

The C5.0 decision tree belongs to the Top-Down Induction of Decision Trees (TDIDT) family of learning system (Quinlan, 1986). Like most machine learning tree scheme algorithms, TDIDT uses the basic strategy of Divide-and-Conquer. This strategy consists of selecting a test for root node; creating branch for each possible outcome of the test; splitting instances into subsets, one for each branch extending from the node; repeat recursively for each branch, using only instances that reach the branch and stop the recursion splits of a branch if all its instances have the same class (Quinlan, 1986, 2014; Witten and Frank, 2005).

A case problem extracted from the "C4.5: Programs for Machine Learning" book, by Ross Quinlan (Quinlan, 1986) will serve as an example for a better understating of how a decision tree works. Table 1 is a small dataset that shows which days a match of tennis was played based on suitable weather conditions.

Table 1: Dataset example for *play tennis*. Extracted from Quinlan, 1986

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

In the above dataset, each day of the table corresponds to a different instance, totaling 14 distinguished instances. Each variable in the dataset is an attribute, which measures different characteristics of the instance, like the weather variables in the problem: outlook, temperature, humidity and wind. The class is the target variable that the decision tree is built to determine; in the above example the classes are "yes" and "no" to play tennis. Figure 3 is the resulting decision tree for the given problem.

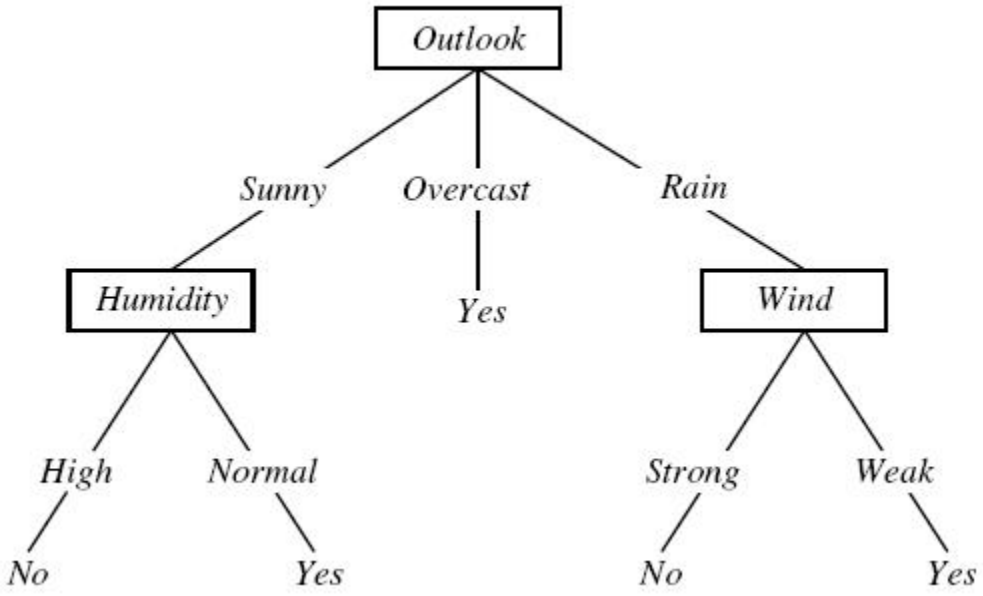

Figure 3: Decision Tree example for *play tennis* (Quinlan, 1986)

A C5.0 decision tree is composed of a root node, decision nodes, branches and leaf nodes (Quinlan, 1986, 2014). The Root Node corresponds to the top node of the tree, represented by the most significant attribute split. In the example, outlook was chosen as the root node. The decision nodes are subsequent nodes represented by attributes, which will further split the remnant data that came from previous splits. Humidity and wind are subsequent decision nodes that additionally divide the rest of the dataset that reached this node in order to properly reach the correct output. The branches are the link between a root or decision node to a subsequent decision or leaf nodes. Branches are always categorized by a nominal or numeric value present under an attribute instance. In the example, sunny, overcast and rain are braches of the root node (outlook), while high and normal and strong and weak are the branches of the splits made by the decision nodes (humidity and wind). The leaf nodes represent the class outcome (target) of the decision tree scheme. “Yes” and “no” are the leaf nodes of the tree.

2.2 Splitting Criteria

Splitting the data consists is done by dividing the values of an attribute in order to best separate the classes. Numeric attributes are split in two branches; one consisting of higher values of a specific number and the other with lower values from that same number. Nominal attributes are split according to the number of different non-numerical values of that specific attribute.

The choice of the attribute and the splitting value criteria is based on the information gain ratio related to that attribute (Quinlan, 2014). The attribute with the highest information gain ratio is chosen to make the decision at a node, the subsequent decision nodes are chosen the same way, based on the leftover data. Information gain is a measurement of the expected reduction in entropy in the data produced by a split (Mitchell, 1997). The decision at each node of the tree is made based on the subset of the data that maximizes the reduction in entropy of the descendent nodes (DeFries and Chan, 2000).

Information Gain is defined as follows (Quinlan, 1986, 2014):

$$Information\,Gain\,(S, A) = Entropy\,(S) - Entropy\,(S, A)$$

Where Information Gain (S,A) is the expected reduction in entropy caused by knowing the value of attribute A (Mitchell, 1997). In other words, it is the information after attribute A is chosen as a test for the training samples to divide into, in order to better split the target classes S, i.e. *play tennis* results. Entropy is the measure of the uncertainty (impurity) associated with a random variable.

The entropy of a collection of classes S, i.e. *play tennis* results, can be defined as follows (Mitchell, 1997; Quinlan, 2014):

$$Entropy(S) = \sum_{i=1}^{n} p(ci)\, log_2\, p(ci)$$

Where *ci* is one of the possible class outcomes of the set S, and *p(ci)* is the proportion of the class over the whole set S. Using the play tennis problem, *c1* and *c2* would be the classes "yes" and "no", respectively, and *p(ci)* would be proportion of each class, i.e. $p(c1) = 9/14$ and $p(c2) = 5/14$. The graph from figure 4 shows the distribution of values of entropy considering a two-class variable. If the sample is completely homogeneous (only one class) the entropy is zero and if the sample is equally divided it has an entropy of one. Using either the graph or the equation, the entropy for S, for the play tennis example, is 0.940

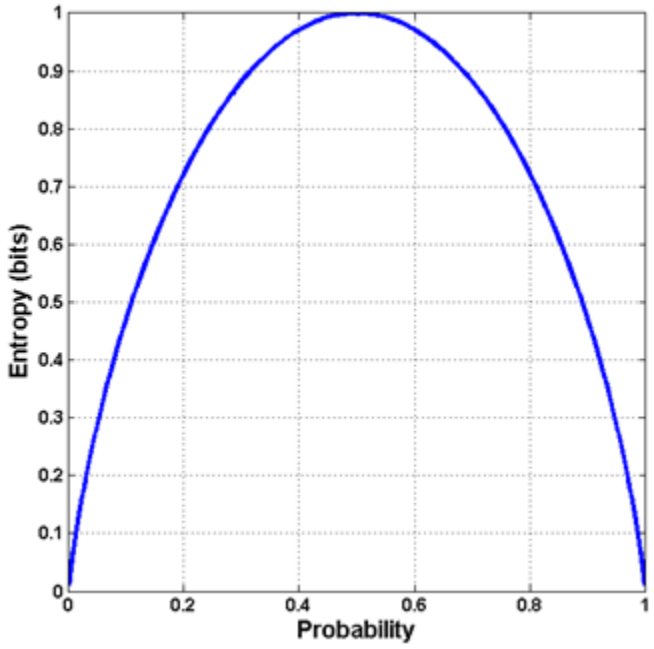

Figure 4: Entropy values distribution for a two class variable (Mitchell, 1997)

The second part of the equation of information gain represents the entropy value of class S distributed through a specific attribute A. The following equation describes Entropy (S,A) (Mitchell, 1997; Quinlan, 2014):

$$Entropy(S, A) = -\sum_{j=1}^{v} \frac{|S_j|}{|S|} * Entropy\,(S_j)$$

Where v is the number of different values for the attribute A, and *Sj* is the subset of S for which attribute A has the value *j*. The expected entropy described by this term is simply the sum of the entropies of each subset Sj, weighted by the fraction of examples that belong to S *(Sj/S)* (Mitchell, 1997).

For example, using the attribute"Wind" as A, S is a collection containing 14 examples, 9 yes and 5 no. Of these 14 examples, 6 are positive ("yes") and 2 are negative ("no") for a *Wind=Weak*, and 3 are positive ("yes") and 3 are negative ("no") for a *Wind = Strong*. The entropy value of the class S distributed through the attribute "Wind" are:

$$Entropy\,(S, Wind) = -\left(\frac{8}{14}\right) * Entropy\,(S_{weak}) - \left(\frac{6}{14}\right) * Entropy\,(S_{strong})$$

Using the graph (figure 4) or the entropy equation, the values for Entropy(S*weak*) Entropy(S*strong*) are 0.811 and 1, respectively.

$$Entropy(S, Wind) = -\left(\frac{8}{14}\right) * 0.811 - \left(\frac{6}{14}\right) * 1.00 = -0.892$$

Using the results of Entropy (S) and Entropy (S—A), the information gain of the attributes can be calculated. The figure 5 was extracted from the "Machine Learning" book, by Tom Mitchel, to better illustrate the information gain calculation.

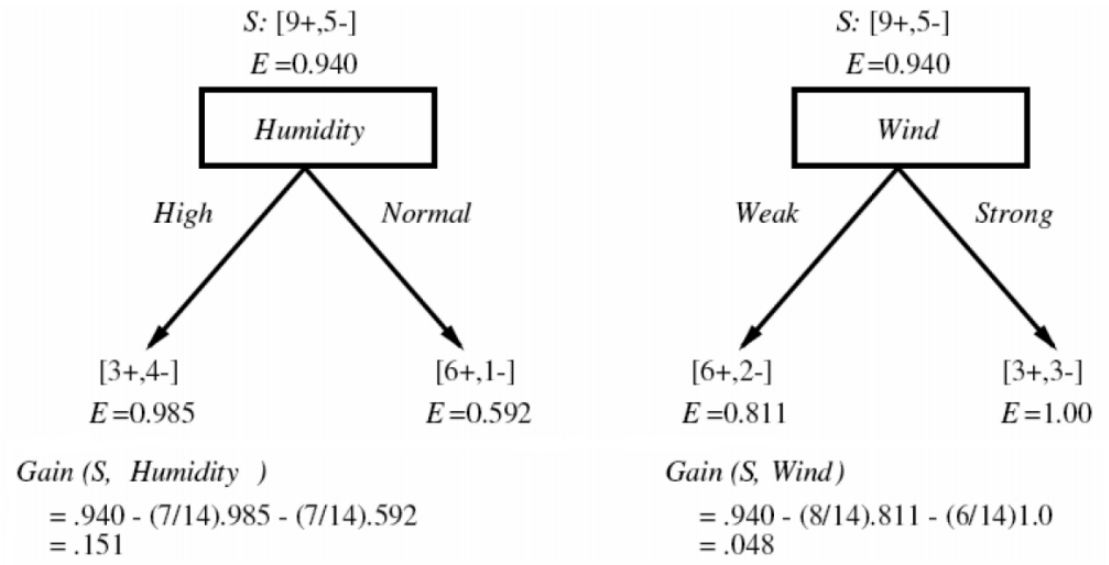

Figure 5: Information Gain calculation for two variables from the *play tennis* problem (Mitchell, 1997)

There is a natural bias in the information gain measure that favors attributes with many values over those with few values (Quinlan, 2014). This affects the tree negatively, because those attributes will lead to a large number of subsets, which grow the tree larger than it is supposed to be. The gain ratio measure penalizes attributes to those attributes by incorporating the split information term, that is sensitive to how broadly and uniformly the attribute splits the data (Mitchell, 1997; Quinlan, 2014). The split information term discourages the selection of attributes with many uniformly distributed values (Mitchell, 1997).

The information gain ratio and the split information equations are defined below (Mitchell, 1997; Quinlan, 2014):

$$Gain\,Ratio\,(S,A) = \frac{Information\,Gain\,(S,A)}{Split\,Information\,(S,A)}$$

$$SplitInformation(S,A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} * log_2\left(\frac{|S_i|}{|S|}\right)$$

Where *S1* through *Sc* are the *c* subsets of examples resulting from partitioning S by the *c* valued attribute A; the split information is actually the entropy of S with respect to the values of attribute A (Mitchell, 1997). For example, using the attribute "Wind" once again as A and S as the collection containing 14 examples of "yes" and "no" to play tennis. Of these 14 examples, 8 are related to "Wind=Weak" and 6 for a "Wind=Strong". The split information value of the class S distributed through the attribute "Wind" are:

$$SplitInformation(S, Wind) = -\frac{8}{14} * log_2\left(\frac{8}{14}\right) + \frac{6}{14} * log_2\left(\frac{6}{14}\right) = 0.985$$

The resulting information gain ratio for the class wind is:

$$Gain\,Ratio\,(S, Wind) = \frac{Information\,Gain\,(S, Wind)}{Split\,Information\,(S, Wind)} = \frac{0.048}{0.985} = 0.049$$

2.3 "Overfitting" and Pruning

The decision tree algorithm grows each branch of the tree just deeply enough to perfectly classify the training examples which can lead to difficulties when there is noise in the data, since fully expanded decision trees often contain unnecessary structure (Quinlan, 2014). This full grown tree might not fit any data as well as it fits the training data, which invalidates the classification purpose. This phenomenon is called "overfitting" and is particularly likely to happen when the number of parameters in the model is large (Quinlan, 2014).

To avoid overfitting the model, it is generally advisable to simplify the trees before they are tested or used for prediction. This simplification of a decision tree is called pruning. The pruning process can reduce significantly the size of the decision tree without losing much of its accuracy. Pruning can be performed in two distinct manners; by stopping the growth of a branch when information becomes unreliable, when there is no statistically significant association between the predictors and the target

class, a method called pre-pruning; or by a later simplification after the tree is built, a method called post-pruning (Witten and Frank, 2005).

The C5.0 algorithm utilizes post-pruning techniques to simplify the final tree. After the tree is built, the model estimates its error rate based on a confidence interval from the training data (Witten and Frank, 2005). The simplification can be achieved by two different operators; subtree replacement and subtree raising (Witten and Frank, 2005). The first consists of replacing a small part of the tree (subtree) with a single leaf node; the second operator consists of deleting an intermediate decision node, replacing it with subsequent nodes (Witten and Frank, 2005). Both methods can drastically reduce the size of the tree, avoiding the overfitting phenomenon.

CHAPTER III

SITE DESCRIPTION

The four sites studied are the Jemez and Santa Fe mountains in New Mexico, the Blue Mountains in Oregon, North Cascade mountain region in Washington and the entire state of Wyoming. All of those sites present differences in size, climatic conditions, vegetation types and elevation range. The following sections will briefly describe these sites, along with visual surface reflectance (Landsat), USGS 2001 land cover, elevation, annual mean temperature and annual precipitation maps.

The following figure displays the location and the size of the four sites in USA.

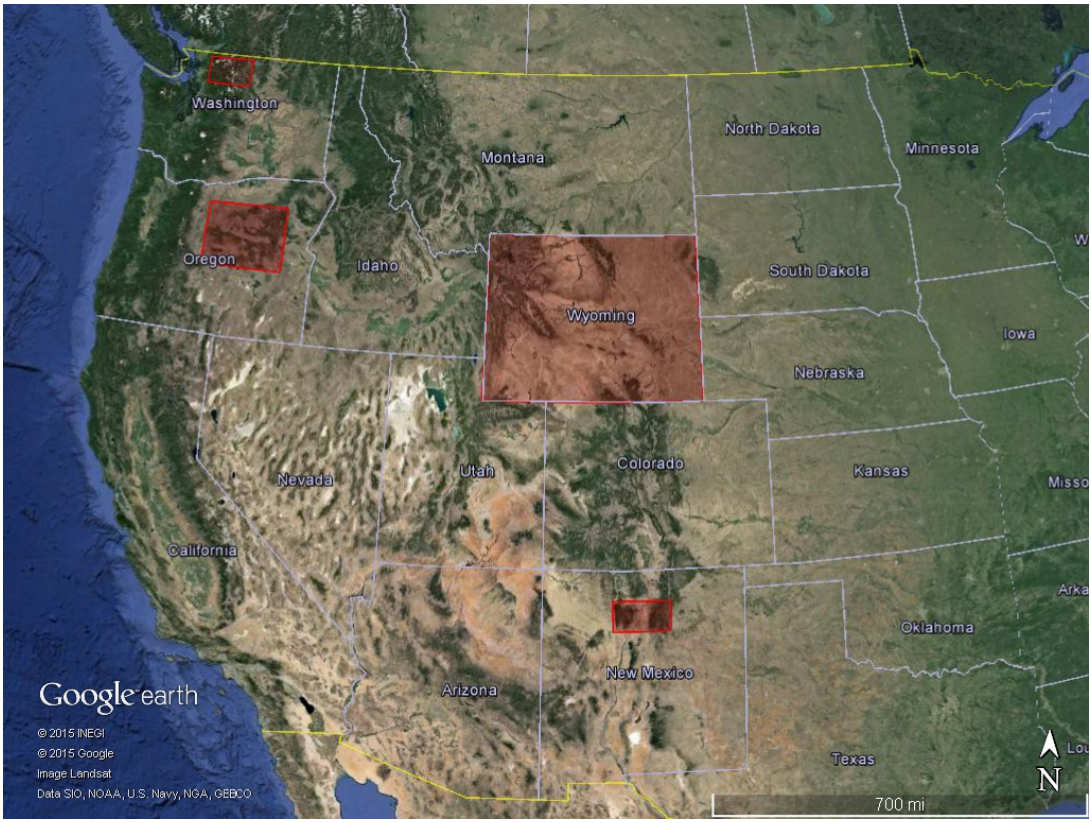

Figure 6: Location and size of all sites in USA

3.1 New Mexico Site

The New Mexico site is a large area, approximately 10,570 km^2, situated at the northern central part of New Mexico State. The site englobes the Jemez Mountains and the San Pedro Mountains, on the west part of the region, and the Santa Fe Mountains, a subrange of the Sangre de Cristo Mountains, on the east part of the site. The Valles Caldera National Preserve and most of the Santa Fe National Forest are also located within the region. Notable cities located within the region are Los Alamos and Santa Fe.

The elevation ranges from 1573 m around the Rio Grande River, situated at the south part of the region, to 3972 m at the peaks of the Santa Fe Mountains on the east part of the area (Gesch et al., 2002). The predominant type of vegetation is evergreen forest, covering over 57 % of the total area (Homer et al., 2004). Major types of trees present are Ponderosa pines, Pinyon-juniper Woodland, Mixed Conifer forest and Spruce-Fir. Other significant vegetation cover includes shrubs, covering 22 % of the area, grassland, 14 % of the area, and deciduous forest, 3 % of the area (Homer et al., 2004).

The soil type in this region consists mainly of Entisols, Inceptisols and Alisols; there is also the presence of an exposed rock formation at some mountain peaks (Soil Survey Staff, 2015).

The annual mean temperature varies from 1.3°C, at higher altitudes, to 12.7°C, at lower altitudes. Annual precipitation varies between 251 mm to 980 mm (Hijmans et al., 2005).

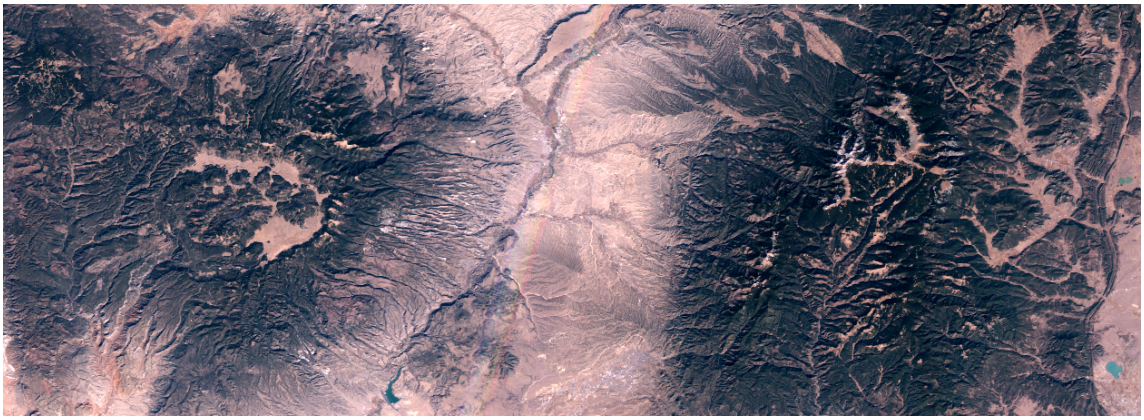

Figure 7: Landsat surface reflectance map of the New Mexico site

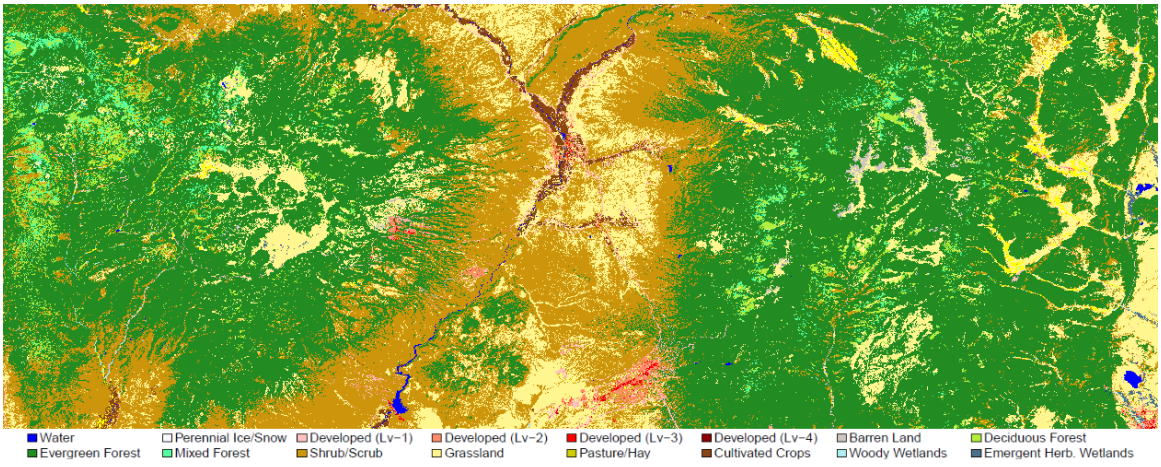

Figure 8: Original 2001 land cover map of the New Mexico site (USGS 2001 NLCD)

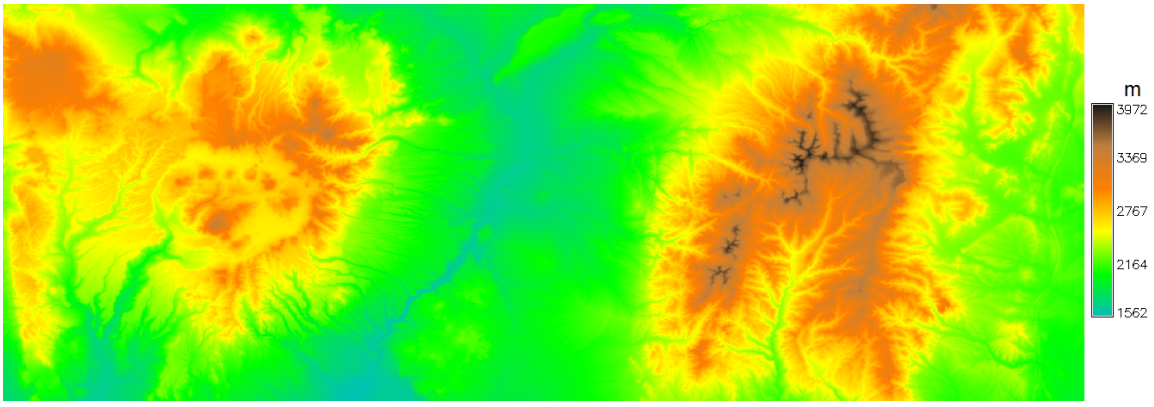

Figure 9: Elevation map of the New Mexico site (USGS NED)

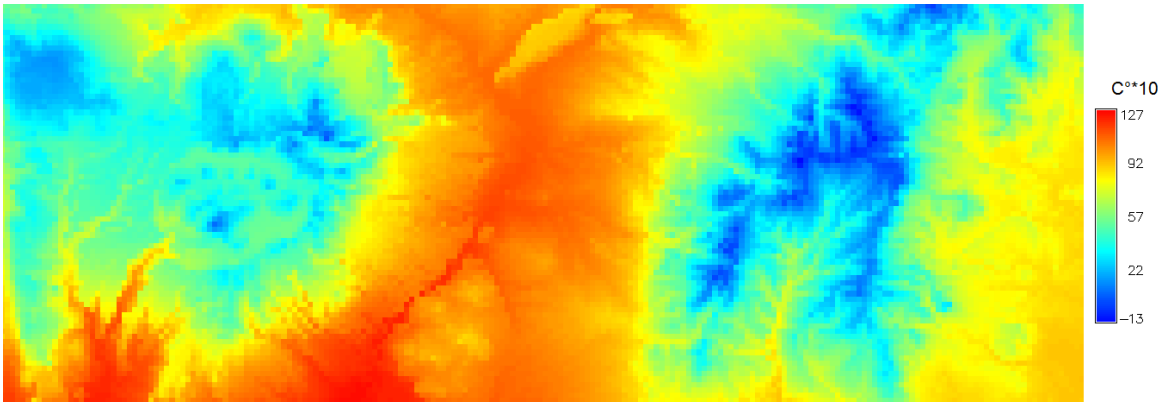

Figure 10: Annual mean temperature map of the New Mexico site (Hijmans et al., 2005)

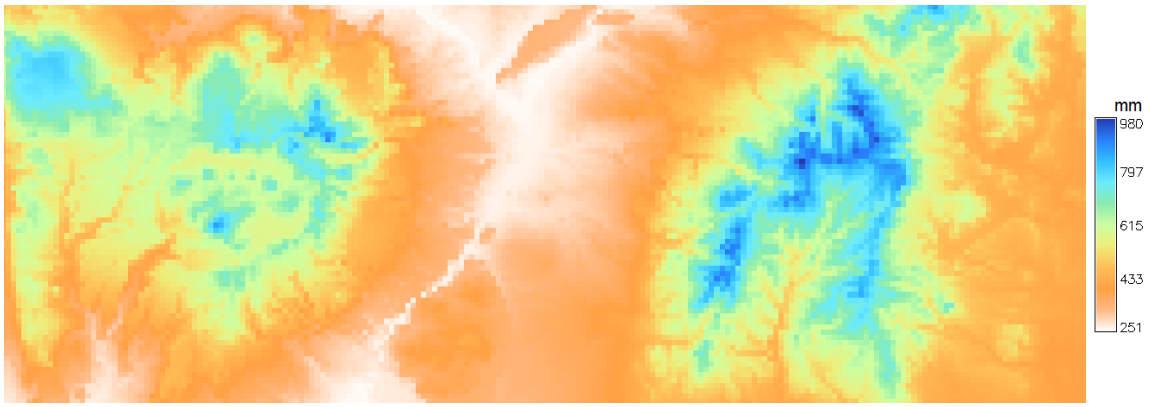

Figure 11: Annual precipitation map of the New Mexico site (Hijmans et al., 2005)

3.2 Oregon Site

The Oregon site is a very large area, approximately 40,070 km^2, situated at the east central part of Oregon State. The site englobes almost the totality of the Blue Mountains region, a large mountain range east of the Cascade Range. The Ochoco National Forest, the Malheur National Forests, the Umatilla National Forest and the Wallowa-Whitman National Forest are all located within this region.

The elevation ranges from 254 m around the Rio Grande River, situated at the south part of the region, to 2758 m at the peaks of the Santa Fe Mountains on the east part of the area. The predominant types of vegetation are shrubs, compound mainly by sagebrush steppe, covering over 59 % of the total area, and evergreen forest, covering 38 % of the site area (Homer et al., 2004). Major types of trees present are Ponderosa Pine, Western Juniper, Mixed Conifer, Mountain Hemlock, Subalpine Fir and Lodgepole Pine (Powell et al., 2007).

The soil type in this region consists mainly of Mollisols and Andisols types (Soil Survey Staff, 2015).

The annual mean temperature varies between 2.1°C, at higher altitudes, to 11.9°C, at lower altitudes. Annual precipitation varies between 217 mm to 825 mm (Hijmans et al., 2005).

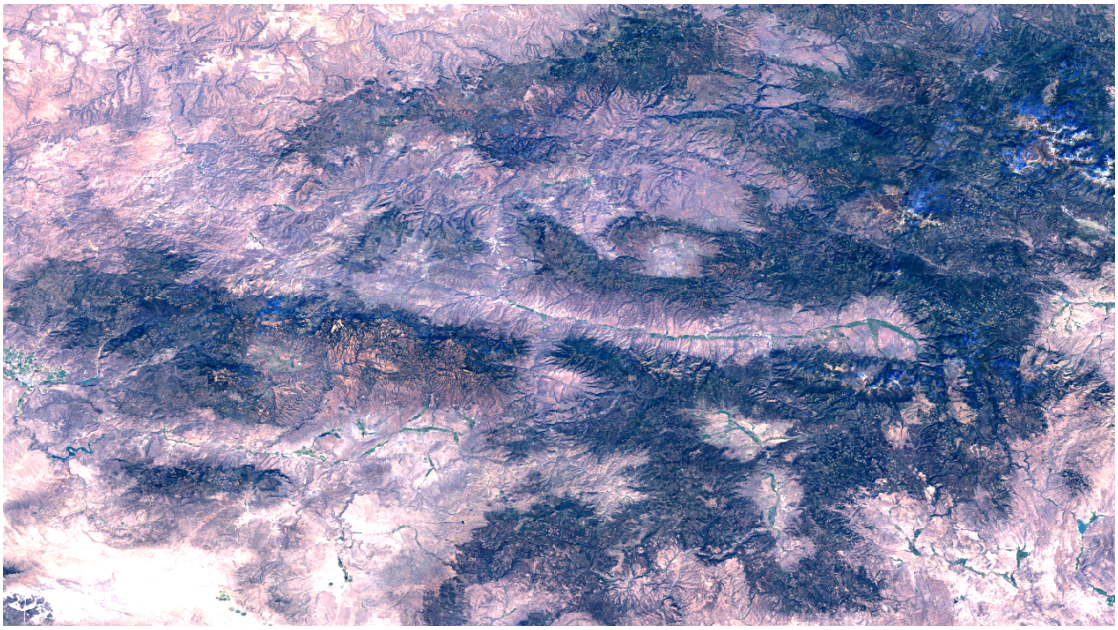

Figure 12: Landsat surface reflectance map of the Oregon site

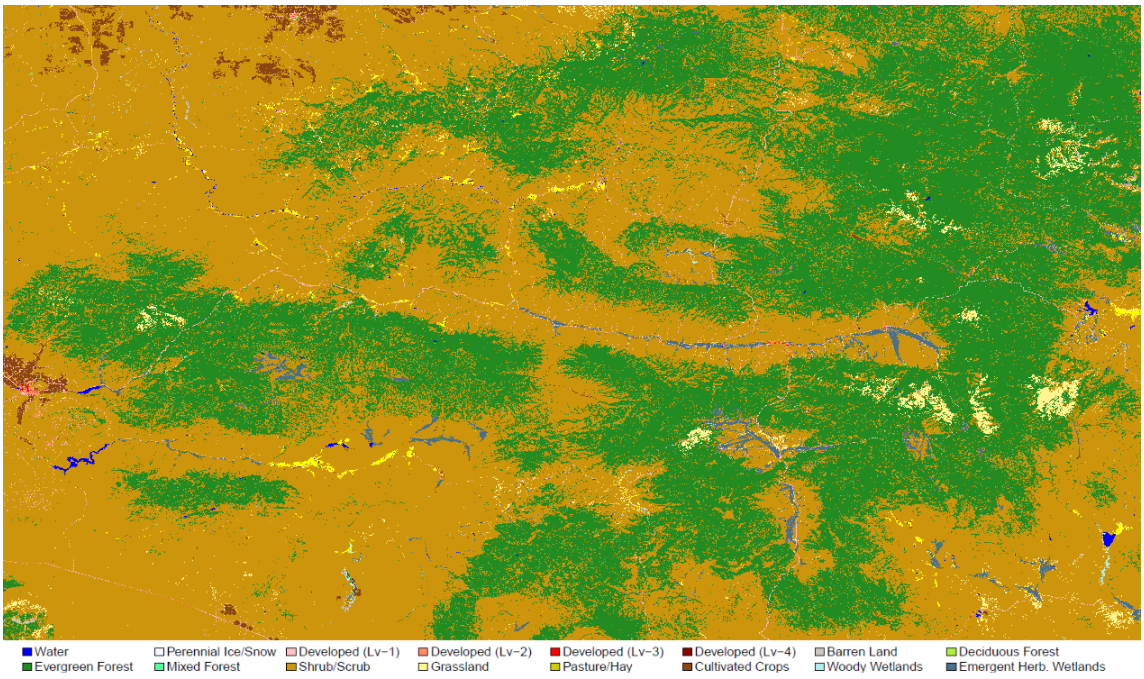

Figure 13: Original 2001 land cover map of the Oregon site (USGS 2001 NLCD)

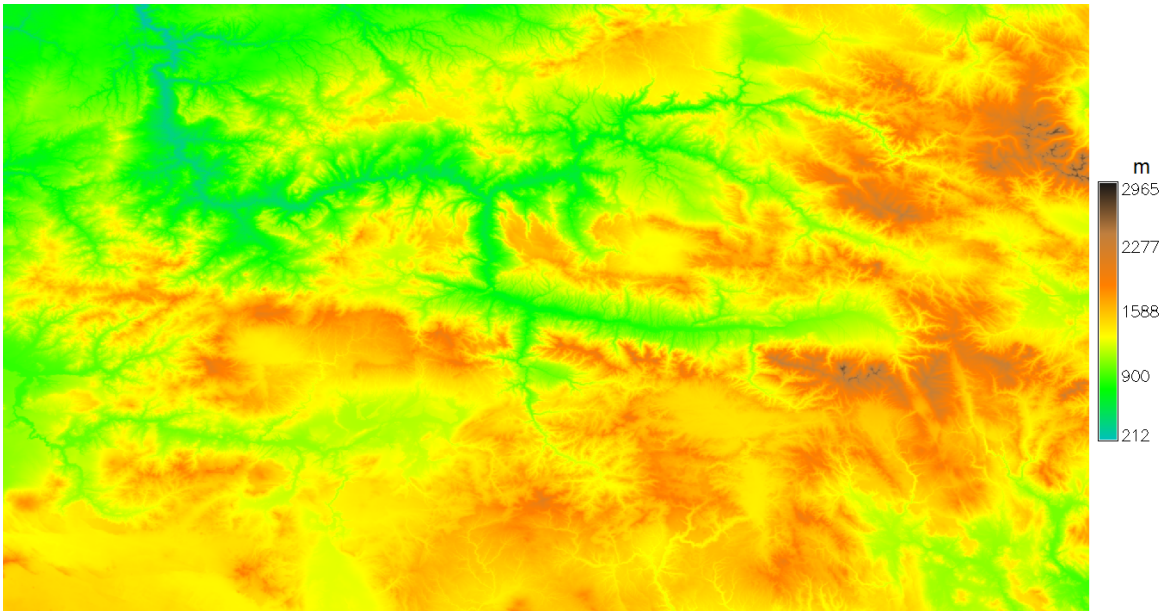

Figure 14: Elevation map of the Oregon site (USGS NED)

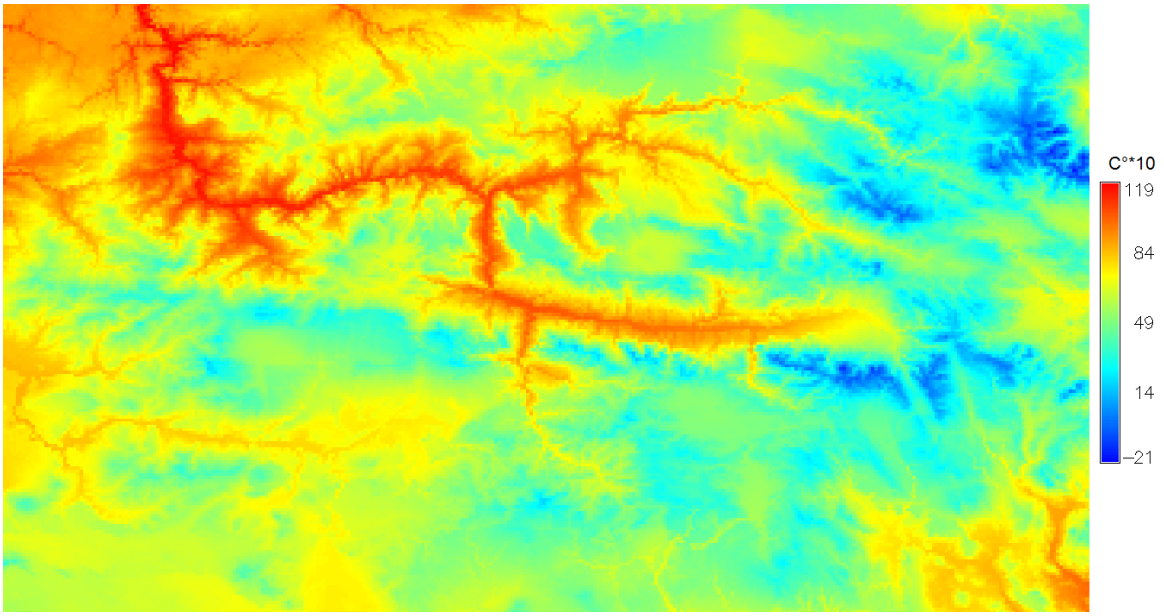

Figure 15: Annual mean temperature map of the Oregon site (Hijmans et al., 2005)

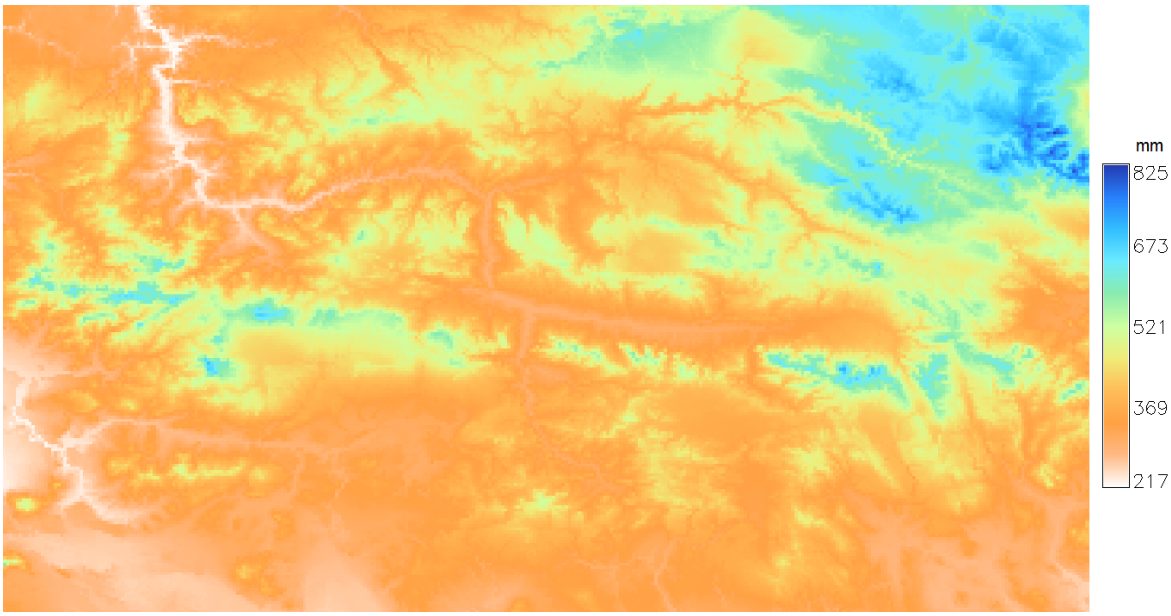

Figure 16: Annual precipitation map of the Oregon site (Hijmans et al., 2005)

3.3 Washington Site

The Washington site is the smallest studied site in this research, with an area of approximately 9,400 km^2, situated at the northwest part of Washington State. The site englobes the North Cascade Mountains, which is a section of the Cascade Mountain Range of western North America. The North Cascades National Park and part of the Mount Baker-Snoqualmie National Forest are also located within the region.

The elevation ranges from 72 m, at the lowest part of the Skagit River, situated at the southwest part of the site, to 3285 m, at the peak of Mount Baker. The predominant type of vegetation is evergreen forest, covering over 61 % of the site area (Homer et al., 2004). Major types of tree present are Western Hemlock, Pacific Silver Fir, Subalpine Mountain Hemlock, Alpine, Subalpine Fir and Douglas Fir (Crawford et al., 2009). Other significant vegetation covers are shrubs, covering 14 % of the territory, grassland, 8 % of the area, and deciduous forest, 1 % of the area (Homer et al., 2004).

The soil type in this region consists mainly of andisols, inceptisols and exposed rock formation (rock outcrops) at higher altitudes of the mountain range (Soil Survey Staff, 2015). The rock outcrops account for almost 8 % of the total surface area of the site.

The annual mean temperature varies from 4.9°C, at higher altitudes, to 10.5 °C, at lower altitudes. Annual precipitation varies between 460 mm, east of the cascades, to 2087 mm at the southwest of the Cascades, close to the coast (Hijmans et al., 2005).

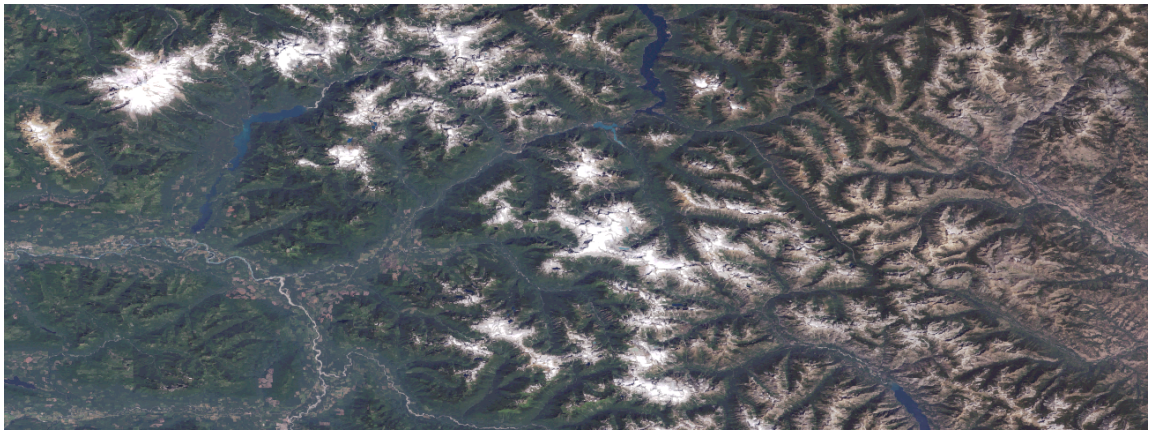

Figure 17: Landsat surface reflectance map of the Washington site

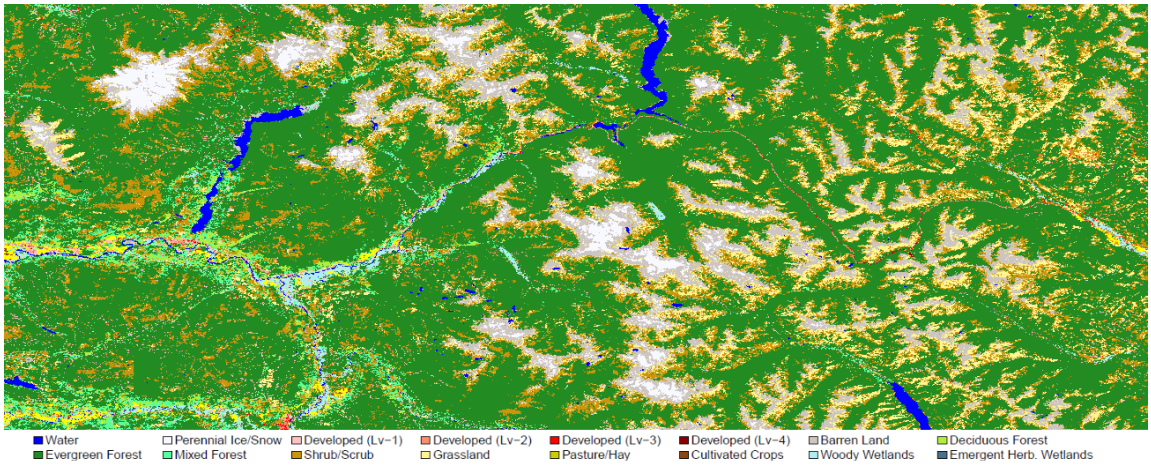

Figure 18: Original 2001 land cover map of the Washington site (USGS 2001 NLCD)

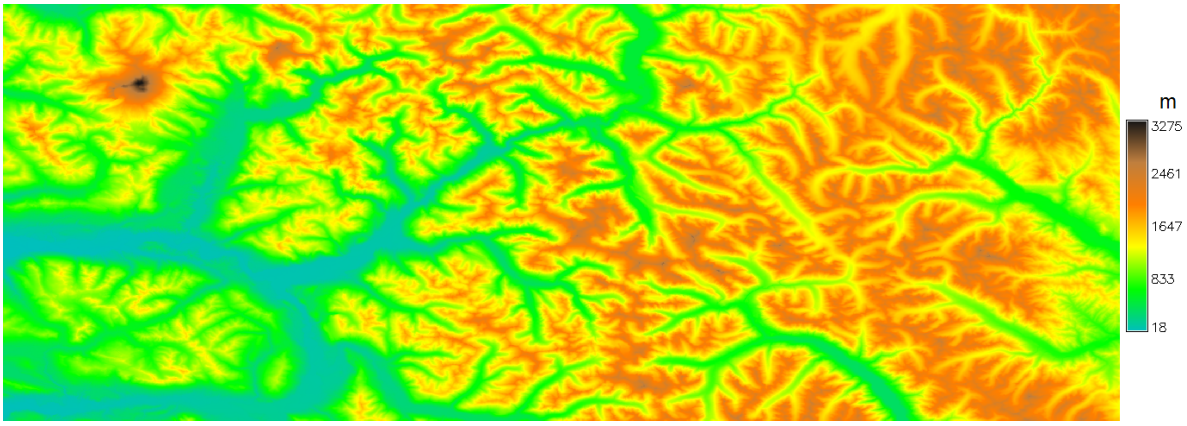

Figure 19: Elevation map of the Washington site (USGS NED)

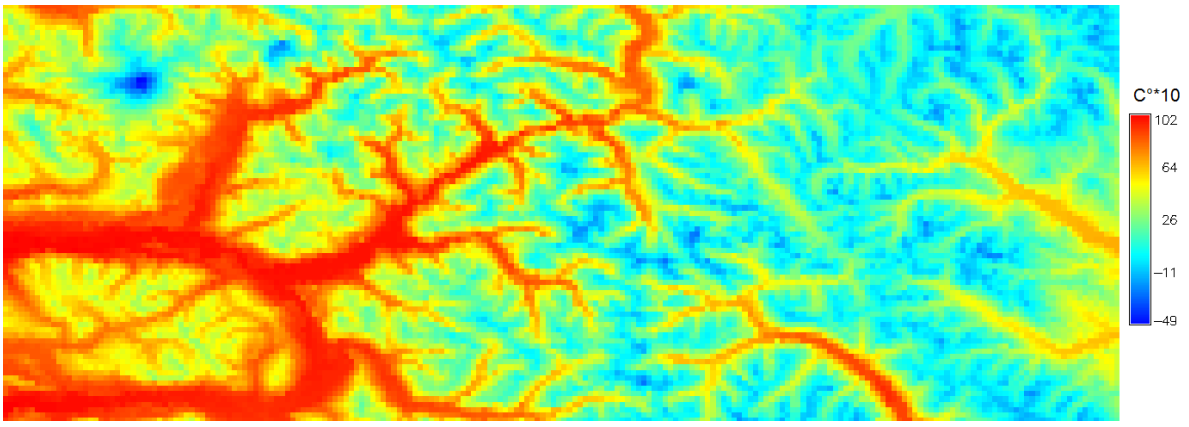

Figure 20: Annual mean temperature map of the Washington site (Hijmans et al., 2005)

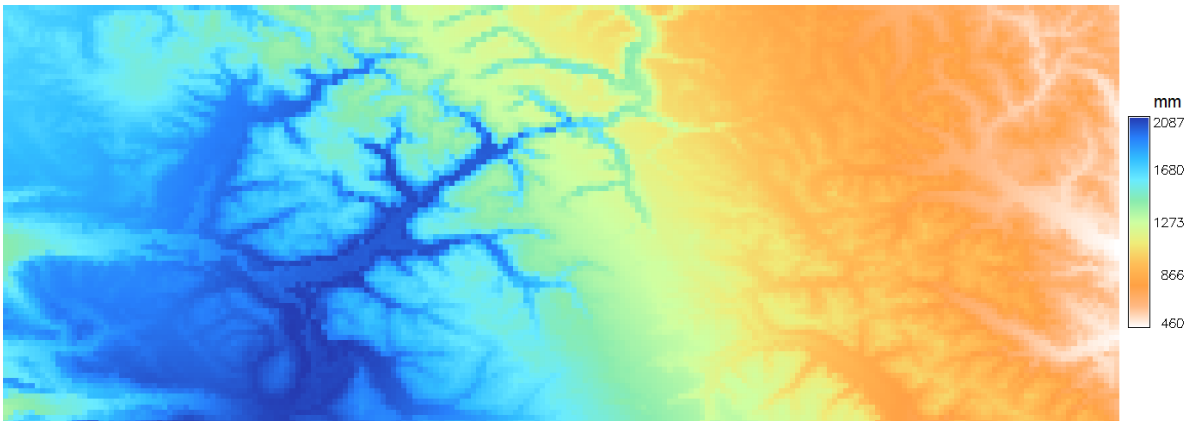

Figure 21: Annual precipitation map of the Washington site (Hijmans et al., 2005)

3.4 Wyoming Site

The Wyoming State is the largest studied site in this work, with an area of 253,348 km^2. While it is the tenth largest state in the United States, it is also the least populated. The state is situated in central west of the country. There are considerable amount of mountains ranges, most of them are part of the Rocky Mountains. In the northwest are the Absaroka, Owl Creek, Gros Ventre, Wind River and the Teton ranges. In the central north are the Big Horn Mountains; in the northeast, the Black Hills; and in the southern region the Laramie, Snowy and Sierra Madre ranges. The Yellowstone National Park, the Grand Teton National Park and several other national forest and natural recreation areas are located in Wyoming State.

The elevation ranges from 945 m, at the Belle Fourche River, situated at the northeast of the State, to 4200 m, at the top of Gannett Peak in the Wind River Mountain Range (NED). The predominant type of vegetation is shrubs, compound mainly by sagebrush steppe and desert shrubs, covering an area over 52 % of the state, and grassland, compound mainly by short mixed-grass prairie, covering 29 % of Wyoming State (Homer et al., 2004). The Evergreen Forest is the third most common vegetation in the State, located almost exclusively in the mountain ranges of the Rocky Mountains at the northwest of the state and in the Bighorn Mountains, at the central north part of Wyoming. Major types of trees present are lodgepole pine, ponderosa pine, spruce-fir, Juniper woodland and Douglas fir (Dorn, 1992).

The dominant soil orders in this region consists of alfisols, aridisols, entisols, gelisols, histosols and molisols and exposed rock formation (rock outcrops) present at the top of the Wind River Mountain Range (Soil Survey Staff, 2015).

The annual mean temperature varies from 6.6°C at higher altitudes of the Wind River and Absaroka mountain ranges at the northwest of the state, to 9.2°C in the

Great Plains on the east. Annual precipitation varies between 172 mm in the Red Desert, located central south of the State, to 847 mm in the region comprised by the Yellowstone National Park (Hijmans et al., 2005).

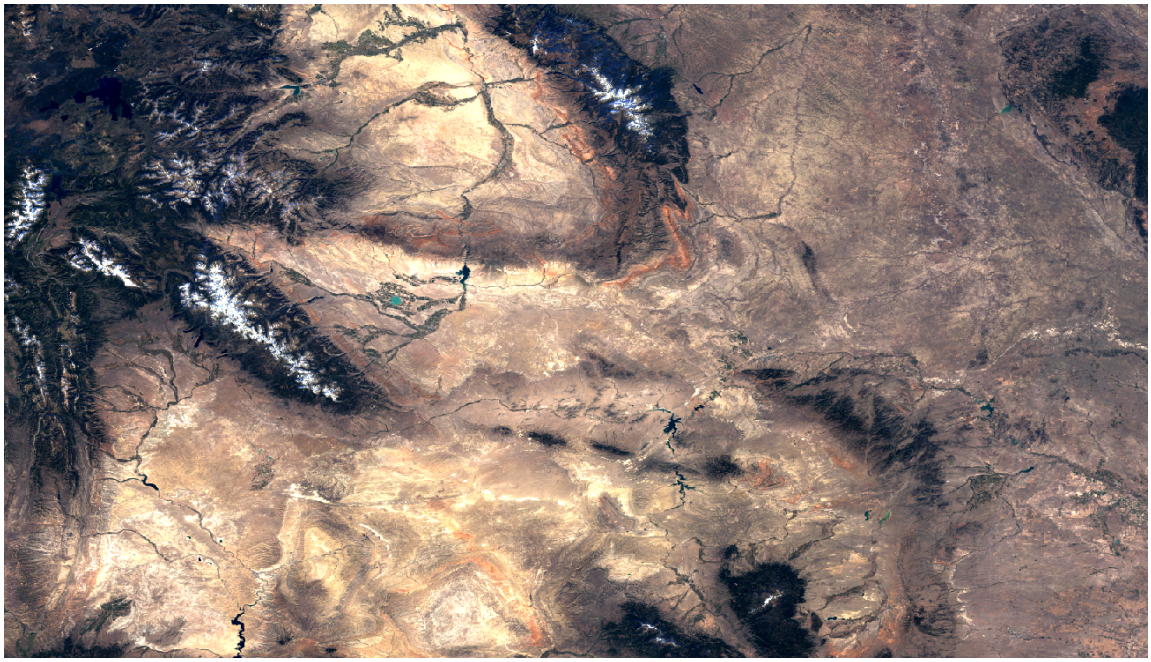

Figure 22: Landsat surface reflectance map of the Wyoming site

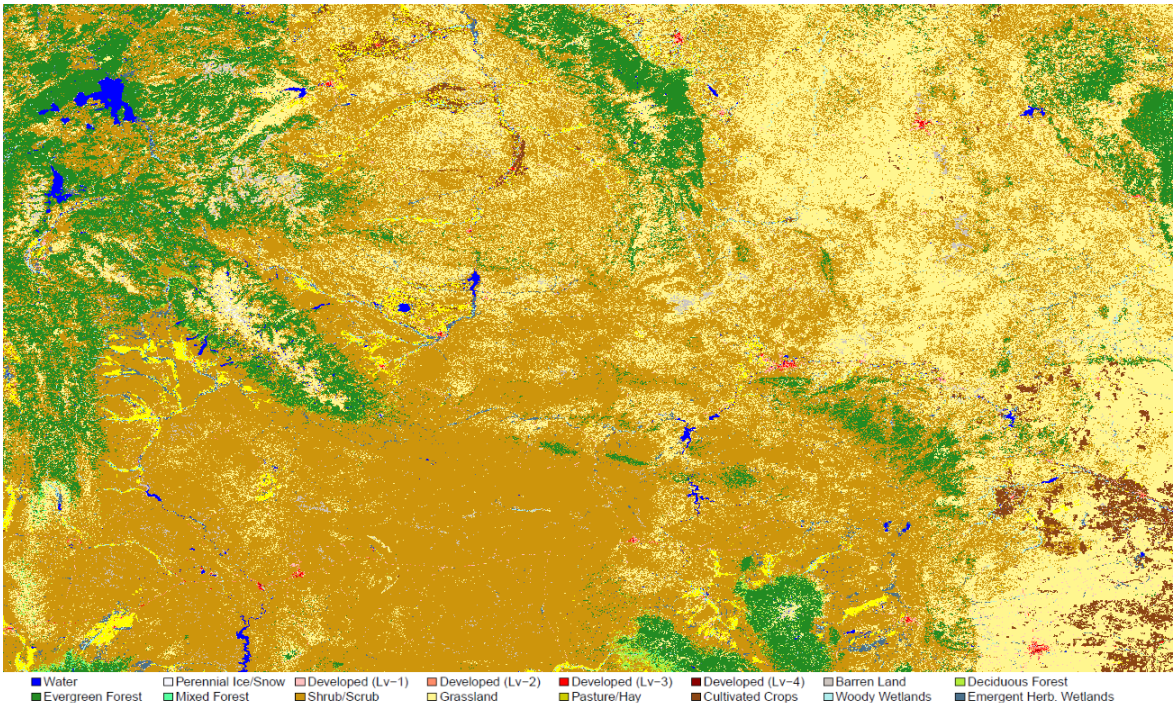

Figure 23: Original 2001 land cover map of the Wyoming site (USGS 2001 NLCD)

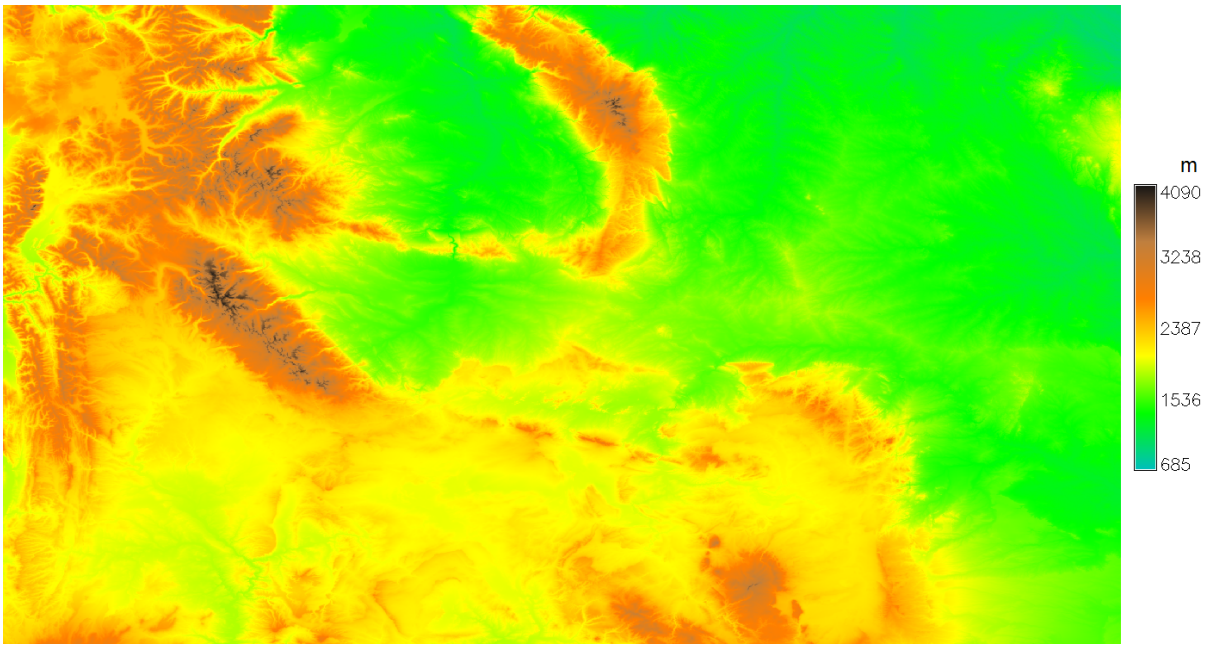

Figure 24: Elevation map of the Wyoming site (USGS NED)

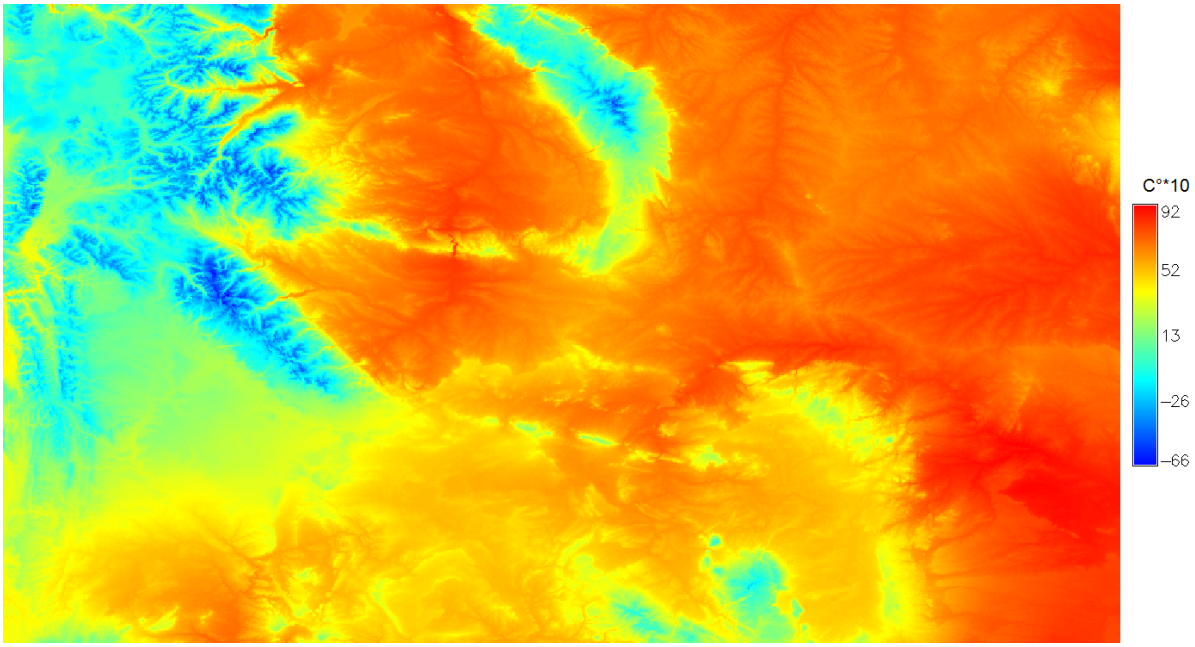

Figure 25: Annual mean temperature map of the Wyoming site (Hijmans et al., 2005)

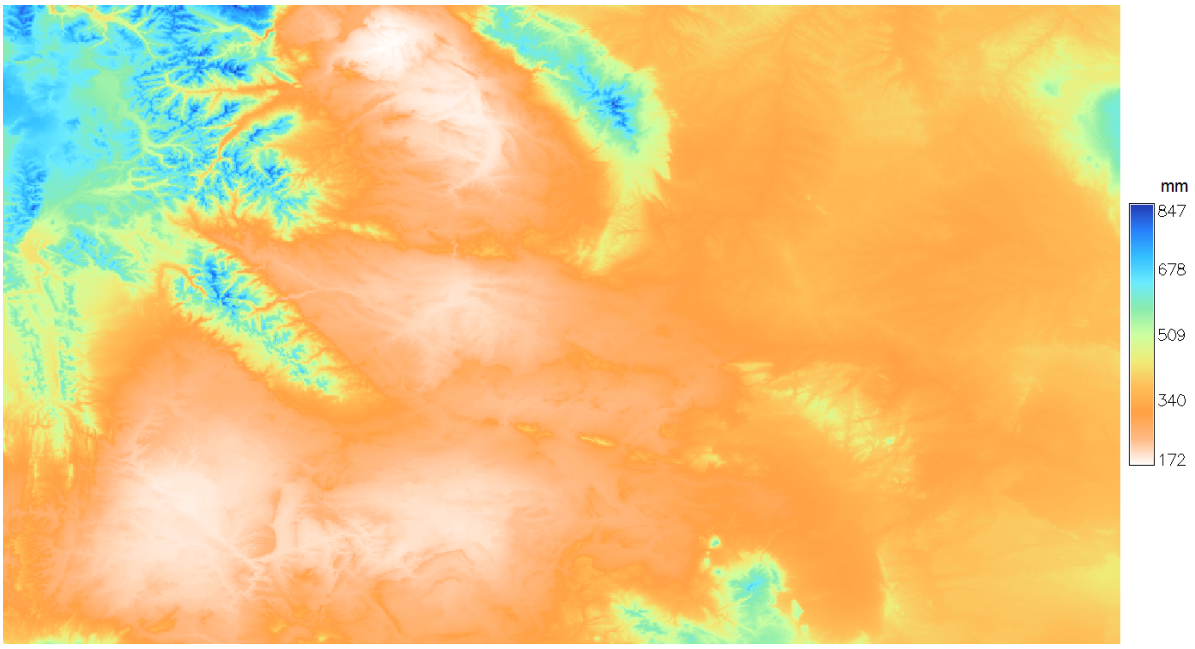

Figure 26: Annual precipitation map of the Wyoming site (Hijmans et al., 2005)

CHAPTER IV

DATA AND SOFTWARE

4.1 Data

Acquiring suitable data is a fundamental step in properly training a machine learning model, since its accuracy and efficiency is totally related to the source data. The datasets, obtained in raster format, were used for training the decision trees were elevation, aspect, slope, land cover and present annual mean temperature and precipitation. Future annual mean temperature and precipitation were used for predicting new results.

4.1.1 Elevation

The elevation data used in this work was derived from the National Elevation Dataset (NED). NED is the major elevation data produced and distributed by the United States Geological Survey (USGS). NED is a multi-resolution dataset that is frequently updated in order to incorporate new or improved data. It provides elevation data for the whole conterminous United States, except for Alaska, Hawaii, and the island territories (Gesch et al., 2002).

NED was chosen because it presents a higher accuracy for the conterminous USA in comparison to other relevant elevation datasets, like the Shuttle Radar Topography Mission (SRTM) by NASA (Gesch et al., 2014).

4.1.2 Slope and Aspect

Slope is the steepness or the degree of incline of a surface. The slope for a particular location is calculated as the maximum rate of change in elevation between that location and its neighbors (Burrough and McDonnell, 2011). The lower the slope value, the flatter the terrain will be; the higher the slope value, the steeper the terrain will be. Slope can be expressed either in degrees or as a percentage (percent rise).

Aspect is the orientation of a slope face, measured clockwise in degrees from 0 to 360, where 0 is north-facing, 90 is east-facing, 180 is south-facing, and 270 is west-facing (Burrough and McDonnell, 2011). Both slope and aspect maps can be calculated and extracted from a digital elevation map.

4.1.3 Land cover

The land cover data, obtained through Google Earth Engine and provided by USGS National Land Cover Database (NLCD), is a 16-class land cover classification scheme that has been applied consistently across the United States at a spatial resolution of 30 meters. It was developed by the Multi-Resolution Land Characteristics (MRLC) consortium consisting of a partnership among several U.S. federal agencies. Currently, there are four different NLCD products available, each one representing the year that the Landsat images were taken and used to produce the national land cover maps.

The 2001 land cover database was chosen to be implemented in this work because of the relative proximity between the data that was used to generate the land cover and the current climate data, comprised from the 1950-2000 period. NLCD 2001 also has an improved classification algorithm compared to the previous land cover database, the NLCD 1992, which was the first one that became available, resulting in

a data with more precise accuracy of spatial boundaries between different land cover classes (Homer et al., 2004).

The 2001 National Land Cover Database was generated from a standardized set of data layers composed mainly by multi-season Landsat 7 imagery and ancillary data, including Digital Elevation Model and its derivatives, including slope, aspect and topography positional index (Homer et al., 2004). Topography positional index is the difference between a cell elevation value and the average elevation of the neighborhood around that cell (Weiss, 2001). One other derivative that was also used for improving land cover classification for developed areas was the per-pixel estimates of percent imperviousness, which corresponds to impenetrable surfaces such as rooftops, roads, or parking lots (developed land) (Yang et al., 2003).

The criteria used for acquiring Landsat images for the development of the 2001 NLCD was based on vegetation phenology and image quality (cloudiness, haze). Optimal time periods for determining distinguished land cover types were identified for each Landsat path/row from which three Landsat scenes were selected, representing early, peak, and late vegetation phenology stages, corresponding to spring, summer and fall seasons, respectively. Those stages are well represented by measuring the vegetation greenness, which is derived from the multi-temporal normalized difference vegetation index (NDVI) data (Yang et al., 2001). Acquiring satellite images from different periods is fundamental to determine and classify different types of vegetation, such as evergreen and deciduous forest. Exceptions to acquiring images outside those determined periods happened only when good quality, usually cloud free scenes, were not available (Yang et al., 2001).

After the Landsat images are collected, they are geometrically corrected (terrain correction) to improve geolocation accuracy (Homer et al., 2004). Later, the image

noise present in some Landsat images, corresponding to the influence of the atmospheric and illumination geometry effects, are normalized and converted to at-satellite reflectance (Homer et al., 2004).

The USGS land cover classification was accomplished using the C5.0 decision tree algorithm. In order to precisely classify the whole USA, the decision tree method requires a considerable amount of training data to build the model, however the amount of data needed for the training is only a small portion of the territory. Various sources were used to provide land cover training data, including reference data for NLCD 1992, high-resolution orthoimagery and field collected points and Forest Inventory Analysis (FIA) plot data. These training data sets were used to map all land cover classes except for the urban classes which were derived from the imperviousness data product (Homer et al., 2007). After the training data is collected and the decision trees are built, the trained model is used to predict land cover classes for the whole USA territory, using all of the Landsat and ancillary data collected. Although the most important data for predicting is the Landsat, the ancillary data (elevation and its sub products) were relevant for the full weighting in the classification process (Homer et al., 2004). The result is a 16 class national land cover map. The classes are: Open Water, Perennial Ice/Snow, four different intensities for developed areas, Barren Land (Rock/Sand/Clay), Deciduous Forest, Evergreen Forest, Mixed Forest, Shrub/Scrub, Grassland/Herbaceous, Pasture/Hay, Cultivated Crops, Woody Wetlands and Emergent Herbaceous Wetlands. Detailed descriptions of the land cover classes are available in the appendix of this thesis.

The decision trees used for training the land cover data had their accuracy assessed by cross-validation and independent data assessment (Homer et al., 2004). Cross-validation can provide relatively accuracy estimates when reference data samples that are statistically valid for both training and accuracy assessment were used (Michie

et al., 1994).

4.1.4 Current Climate Data

Climate data for current conditions was obtained from the WorldClim database, available online at www.worldclim.org. WorldClim was developed by Robert J. Hijmans, Susan Cameron, and Juan Parra, at the Museum of Vertebrate Zoology, University of California, Berkeley, in collaboration with Peter Jones and Andrew Jarvis (CIAT), and with Karen Richardson (Rainforest CRC) (Hijmans et al., 2015).

The WorldClim data consist of different climate layers that cover the all of the global land areas except Antarctica. The climate surface data was generated through interpolation of monthly climate data measured at weather stations from a large number of global, regional, national, and local sources, for the 1950–2000 period, on a 30 arc-second resolution grid (equivalent to 0.86 km, at the equator) (Hijmans et al., 2005). The climate variables that were used in the WorldClim data were monthly precipitation and mean, minimum, and maximum monthly temperature. There are also other variables available, like mean annual temperature, annual precipitation, annual range in temperature and precipitation, mean temperature of the coldest and warmest month, and precipitation of the wet and dry quarters (Hijmans et al., 2005).

Climate data was assembled from numerous sources containing weather station data, including Global Historical Climatology Network (GHCN), which corresponded for the major part of the input data, the Food and Agriculture Organization of the United Nations (FAO), the World Meteorological Organization (WMO), the International Center for Tropical Agriculture (CIAT), R-HYdronet, and other minor

databases for some specific countries (Hijmans et al., 2005). All stations were rigorously checked for accurately reported location, elevation and data consistency. Only weather stations that had at least 10 years of monthly data were considered for the dataset (Hijmans et al., 2005).

The collected data was interpolated with the ANISPLIN software. This software uses every station as a data point and fits thin plate smoothing splines (usually second- or third-order polynomials) through station data in three independent variables: latitude, longitude, and elevation (Hutchinson, 1995; Hijmans et al., 2005). The elevation data used for this project was the Shuttle Radar Topography Mission (SRTM) with an aggregated resolution of 30 arc-seconds. The interpolated data, associated with elevation data, resulted in the climate surfaces.

Two variables from the WorldClim database were used in this work to represent current climate conditions; the mean annual temperature and annual precipitation; both with a 1000 m resolution.

4.1.5 Future Climate Data

Future climate data was generated using Model E2 version of the Goddard Institute for Space Studies (GISS) General Circulation Model (GCMs) for the target year of 2070 (average for 2061-2080 period), downscaled to a 30 arc sec (1 km) resolution and bias corrected for the WorldClim database (Hijmans et al., 2015). The data can be accessed online at www.worldclim.org

General Circulation Models (GCMs) are advanced tools for simulating the response of the global climate system to increasing greenhouse gas concentrations; represented by physical processes in the atmosphere, ocean and land surface. GCMs have the

potential to provide consistent estimates of regional climate change for distinguished past and future scenarios (Carter et al., 2007). These models simulate weather in different layers of the atmosphere for small time steps and they are numerically complex.

The GISS-E2-R model uses a three dimensional grid over the globe, with a horizontal resolution of 2°by 2.5°(around 220 km by 280 km, at the equator) and 40 vertical layers in the atmosphere, with the model top near the stratopause at a height around 60 km (Schmidt et al., 2014). Coupled with the atmospheric model is the Russell ocean model, with a horizontal resolution of 1.25°by 1°(around 140 km by 110 km, at the equator), and 32 vertical levels with finer vertical resolution in the top 100 m (Hansen et al., 2007).

Both anthropogenic and natural forcing agents were included as input variables for the GISS-E2-R simulation runs. Values for those variables, obtained accordantly for each of the four RCP scenarios, were used as input data to simulate future climate data. Anthropogenic forcing variables include: time-varying well-mixed greenhouse gases emissions (CO2, CH4 and N2O), ozone, tropospheric aerosols (sulfates, nitrates, black carbon and organic carbon), stratospheric water vapor from methane oxidation, a parameterized indirect effect of aerosols on clouds, soot effect on snow and ice albedos, and anthropogenic land use changes (Schmidt et al., 2014). Natural forcing agents include: volcanic aerosols, solar irradiance, and Earth orbital parameters (Schmidt et al., 2014).

Using the variables mentioned above, the GISS model output global and regional mean surface air temperature and precipitation. Relative to the 1996-2005 period mean temperature in the historical simulations, simulated global warming ranges from 0.6°C to 3.4°C by 2100 (Nazarenko et al., 2015). For both intermediate RCP 4.5 and

RCP 6.0 scenarios, the warming of the global mean surface air temperature increases by 1.6°C and 2.3°C, respectively by 2100. In the RCP 2.6 scenario, the warming peaks at 2050, where temperatures increases around 0.8 C, and then decreases to 0.6°C by 2100 reflecting the peak and subsequent decline of the radiative forcing. Lastly, the warming in the RCP 8.5, the worst case scenario, reaches 3.4°C by 2100 (Nazarenko et al., 2015). Besides outputting global surface air temperature and precipitation distribution, the GISS E2 model can also generate results for cloud cover, sea ices changes, ocean temperature change and sea level change (Nazarenko et al., 2015).

The resulting climate model outputs for mean temperature and precipitation distribution from the GISS-E2-R model are at a coarse resolution, and this the lower resolution is not compatible with the rest of the data collected, which has a much higher spatial resolution. In order for the data to be suitable, it was downscaled to a resolution of 30 arc sec (known as 1 km resolution) to match the climate surface maps for current conditions from WorldClim. The first step of the downscaling process is computing the difference between the output of the GISS-E2-R for a specific weather variable run for the baseline years (current climate conditions) and for the target years (future climate conditions) (Hijmans et al., 2015). Later, this computed difference is interpolated to a 30 arc sec resolution grid. This higher resolution difference surface layer is applied over the current climate data period, so the resulting future climate map is bias corrected in relation to the present data (Hijmans et al., 2015).

4.2 Software

Two different software products, GRASS GIS and R, and one online environment, Google Earth Engine, were used in order to properly acquire and process data and to train the model.

4.2.1 Google Earth Engine

Google Earth Engine is an online environment monitoring platform for environmental data analysis. It stores over 40 years of historical and current global satellite data and allows high-performance tools to analyze and interpret this information that can then be visualized on a map. The platform was developed by Google, in partnership with Carnegie Mellon University, NASA and the United States Geological Survey (Gorelick, 2012).

The elevation, aspect, slope and land cover raster data were obtained from Google Earth Engine database.

4.2.2 GRASS GIS

GRASS GIS (Geographic Resources Analysis Support System) is a free and open source Geographic Information System (GIS) software used for geospatial data management, analysis and modeling, image processing, graphics and maps production and visualization (Neteler and Mitasova, 2002). It contains several modules to render maps and images; manipulate raster, and vector data; process multispectral image data. GRASS GIS was originally developed by the U.S. Army Construction Engineering Research Laboratories, a branch of the US Army Corp of Engineers, as a tool for land and environmental management and planning by the military (Neteler and Mitasova, 2002).

GRASS GIS was used as an interface for handling all necessary processes related to importing, correcting and exporting raster data.

4.2.3 R and RStudio

R is an open source programming language and software environment for statistical computing and graphics. It presents many functionalities, such as linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc. (Ihaka and Gentleman, 1996). R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand.

RStudio is an integrated development environment for R programming language. It includes many implementations to facilitate the use of R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management (Racine, 2012).

RStudio was the software used to develop R codes to carry all necessary computing processes related to the training and prediction of the C5.0 decision.

CHAPTER V

METHODOLOGY

5.1 Data Acquisition

The elevation raster data (National Elevation Dataset) and the land cover raster data (2001 NCLD), were both provided by USGS and obtained through Google Earth Engine. Slope and aspect raster data were derived from the original elevation map, with a resolution of 1/3 arc-second, through the use of the Slope and Aspect calculation tool from Google Earth Engine.The resolution chosen for each location was carefully changed so the final data would have between 500,000 to 1,000,000 data cells (data instances). If the resolution was maintained as 30 meters the resulting amount of data instances would be extremely large, which would incapacitate the decision tree building process. For the New Mexico and Washington sites, the base resolution was 150 meters. For the Oregon and Wyoming sites, the base resolutions were 250 and 600 meters, respectively. The 1984 version of the World Geodetic System, commonly referred as WGS 84, was chosen as the default coordinate system.

Current and future climate were obtained from the WorldClim database. In total, five different raster datasets were used for each variable, annual mean temperature and annual precipitation. The data is comprised in current conditions, for the 1950–2000 period, and four different future scenarios for the 2070 year. A resolution of 1000 meter was used for all the sites, and the WGS 84 was also chosen as the default coordinate.

The difference in raster resolution between the climate data and the topography and land cover data did not affect the decision tree training.

5.2 Data Preprocessing

All the collected raster datasets are imported into different project frameworks in GRASS GIS software, one for each site location. The imported data is then checked to ensure there are no missing values due to the fact that occasionally some raster data might present missing data around the edges of the map. In the event there is any missing data for any of the raster datasets, a new slightly smaller region is defined, subsequently excluding undesirable areas. After all corrections are completed, the data is extracted into a single comma-separated value (csv) file.

5.3 C5.0 Classification Tree

After all data is pre-processed in GRASS GIS and exported into a csv file, it is finally imported into the R work environment. A code script was developed in R to carry over all necessary computing processes related to the training of the C5.0 decision tree and its use, to classify new land cover for the studied sites. Pre-established code packages used to generate the results were: the C5.0 package and the caret package (Kuhn, 2008), both used for training the decision tree model and applying it to predict new results, and the FSelector package (Romanski et al., 2013), used to determine the attribute importance for training the classification tree model.

Primarily, the imported csv file, which contains all attributes grouped together in a single data frame, is split into different data frames, one for the training data (present climate) and four for the prediction data, one for each of the RCPs climate scenarios. The training data attributes are: elevation, aspect, slope, present mean annual temperature, present annual precipitation and the original land cover classification

by USGS (target attribute). Prediction data attributes are: elevation, aspect, slope, future mean annual temperature and future annual precipitation.

With the data sorted, the C5.0 decision tree is built using all of the training data. The first step to properly validate any machine learning algorithm is to define a testing control. The cross-validation method was chosen to be the testing control because it is one of the most commonly used methods employed in the machine learning science field. A 10 fold cross-validation is set so that, during the decision tree training, the dataset is randomly divided into approximately 10 equal parts, where 9 of the pieces are used for training, and the last piece is used for testing. The cross-validation process is then repeated 9 more times, so that each of the 10 subsamples is used exactly once as the validation data. At the end, the model outputs the average error for all 10 folds. If the accuracy of the cross-validation test is significantly lower than the accuracy of the test using all of the training data, without splitting in different segments, it means that the decision tree model is "overffited" for the training data.

The second step is to define specific parameters for the C5.0 algorithm that will control how the tree will be built and pruned.Important pruning parameters, like the confidence factor and the minimum number of sample outcomes from a split (i.e. minimum number of results possible in a leaf node) were, respectively, set to 0.25 (the default value) and 1. Other relevant C5.0 parameters were either left with default values or turned off, like the boosting option. With all parameters set, the classification tree is built using the training data as the predictors and the land cover classes as the targets on the leaf nodes.

The resulting C5.0 classification tree is then used to predict the four future land cover values by separately feeding the RCPs future climate change scenarios data frames into it (temperature and precipitation). The original current climate data,

the training data minus the land cover classes, is also fed into the tree, so that the predicted land cover for the whole site can be compared to the original land cover to access the accuracy of the decision tree.

The results obtained from the decision tree model for land cover prediction were imported back into GRASS GIS with the aim to create the new land cover surface maps for future and current climate conditions, using the previously saved coordinates. Land cover class distribution for all the results are also extracted as a data frame in order to account for the land cover changes across the different climate scenarios and to compare those changes to the current trained land cover map. The trained land cover for current conditions was used as the base data for comparison to the future land cover results, instead of the original land cover.

Aside from generating new results for future land cover maps, the R code was also set up to produce a table with the importance of each attribute (elevation, aspect, slope, mean annual temperature or annual precipitation) to build the C5.0 decision classification tree. The importance is measured based on the information gain ratio of each attribute in relation to the target of the tree, the land cover. This is a reliable method to estimate the importance because the information gain ratio is also the method used by the C5.0 algorithm to split data for the tree.

CHAPTER VI

RESULTS

6.1 Annual mean temperature and precipitation change

The annual mean temperature and annual precipitation projections for the present climate conditions (normal of 1950-2000 period) and the four climate scenarios for the year of 2070 (2061-2080 period), spatially averaged for each location, are displayed below:

Table 2: Annual Mean Temperature (°C)

Site	**Current**	**RCP 2.6**	**RCP 4.5**	**RCP 6.0**	**RCP 8.5**
New Mexico	6.9	8	9.3	9.2	10.2
Oregon	6	7	7.6	7.7	8.2
Washington	3.8	4.9	5.6	5.8	6.2
Wyoming	4.7	5.7	6.8	6.9	7.8

Table 3: Annual Precipitation (mm)

Site	**Original**	**RCP 2.6**	**RCP 4.5**	**RCP 6.0**	**RCP 8.5**
New Mexico	492	498	477	491	450
Oregon	411	418	417	423	412
Washington	1335	1346	1375	1351	1345
Wyoming	356	361	365	376	349

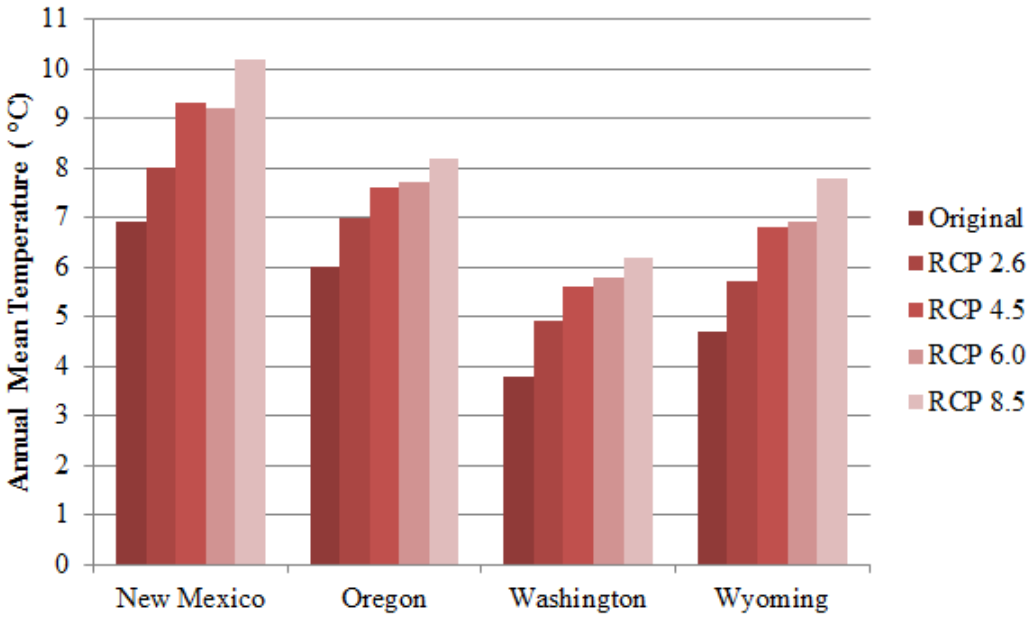

Figure 27: Annual Mean Temperature (°C)

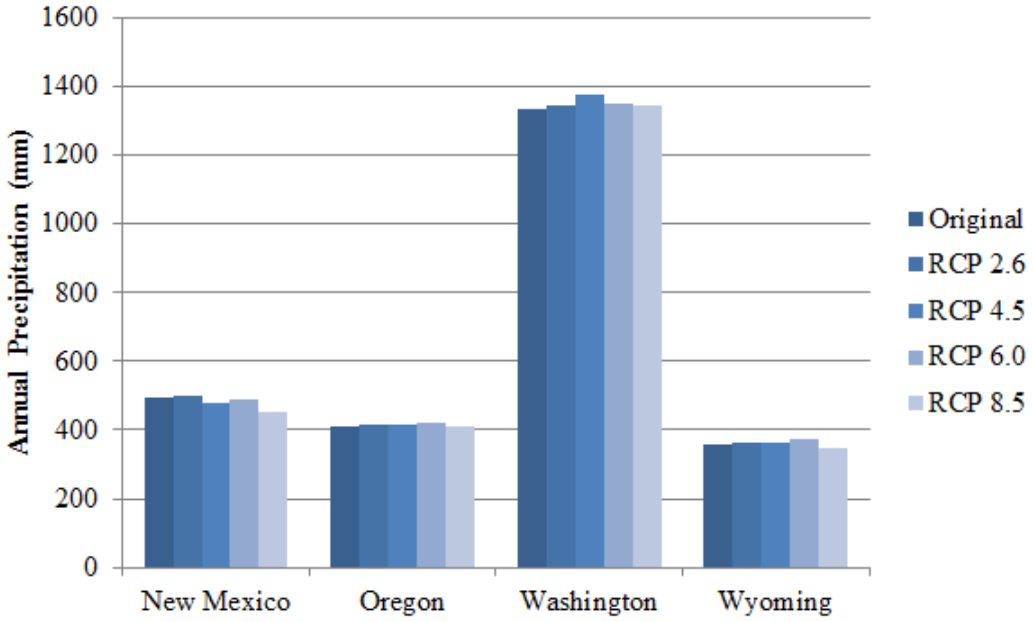

Figure 28: Annual Precipitation (mm)

It's clear that the surface air temperature is higher for all future climate scenarios compared to current climate conditions. The average increase in the annual mean temperature for all four sites are, in order of the RCP scenarios, 1.05°C, 1.97°C, 2.05°C and 2.75°C. Projections for future precipitation are less alarming. The average changes in the annual precipitation for all four sites are, in order of the RCP scenarios, + 7.2 mm, + 10 mm, + 11.7 mm and - 9.5 mm.

The annual mean temperature and annual precipitation maps of each site for the present climate conditions and the four climate scenarios are displayed below:

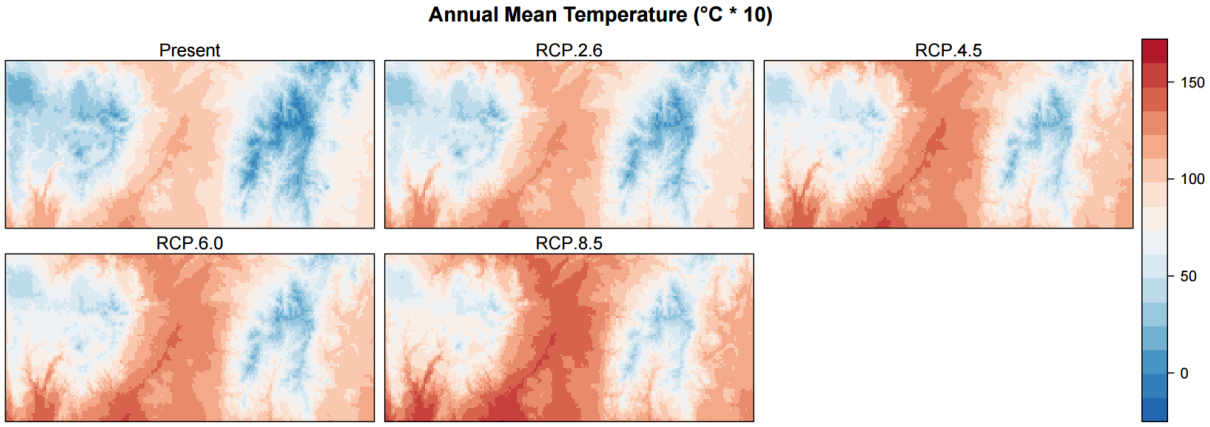

Figure 29: New Mexico annual mean temperature change

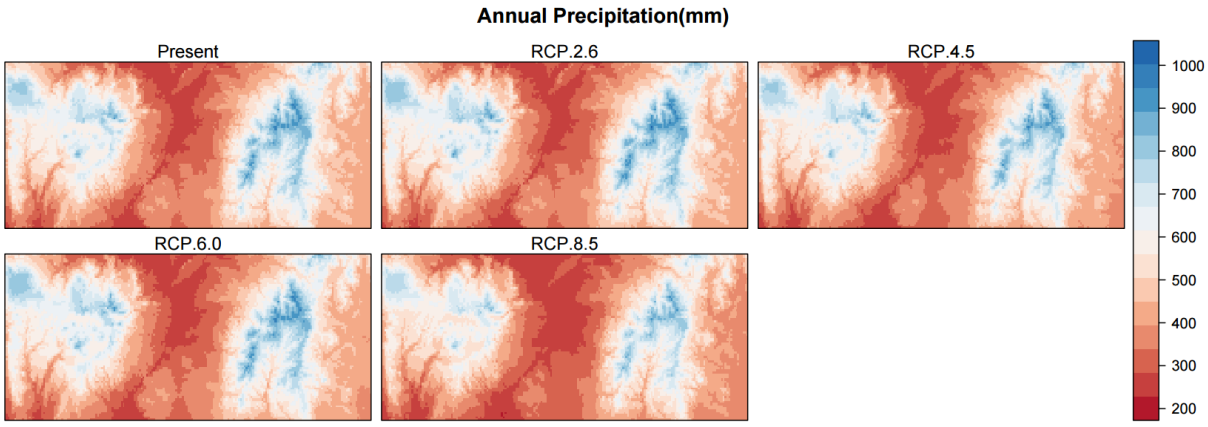

Figure 30: New Mexico annual precipitation change

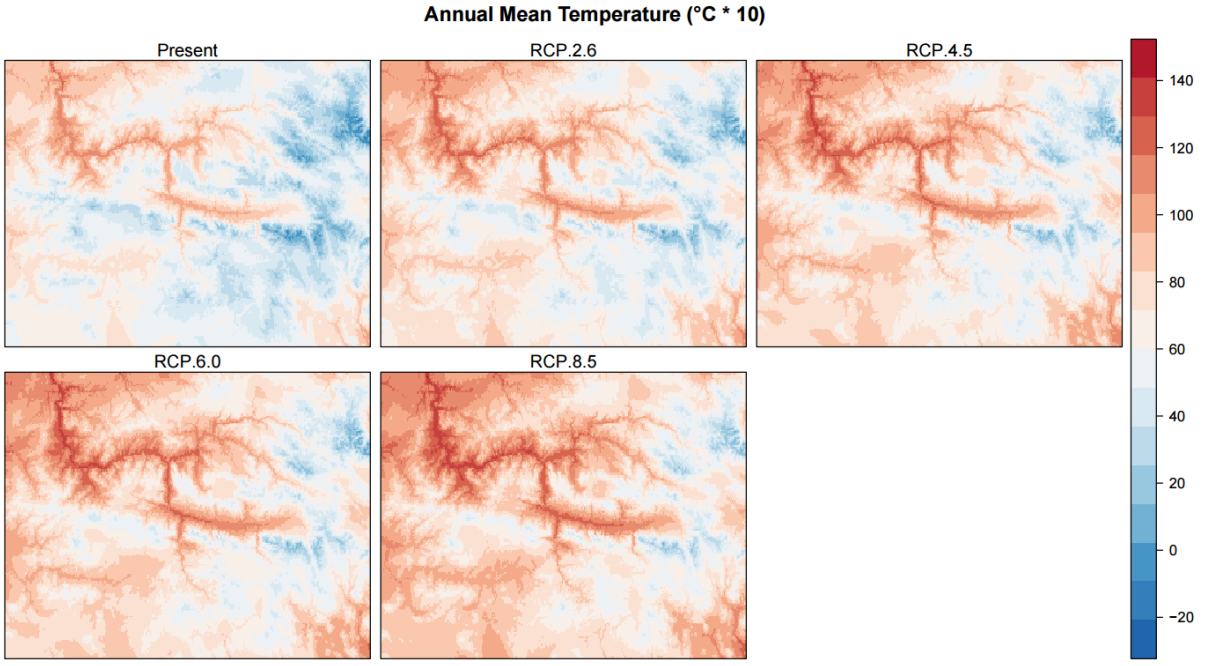

Figure 31: Oregon annual mean temperature change

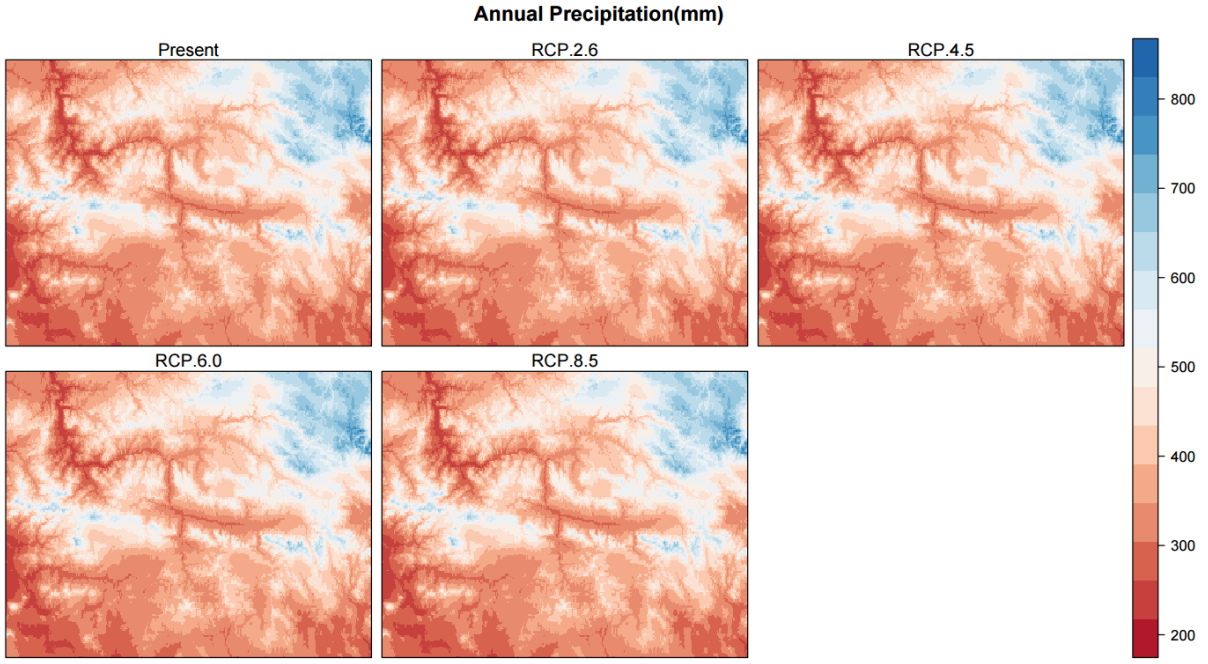

Figure 32: Oregon annual precipitation change

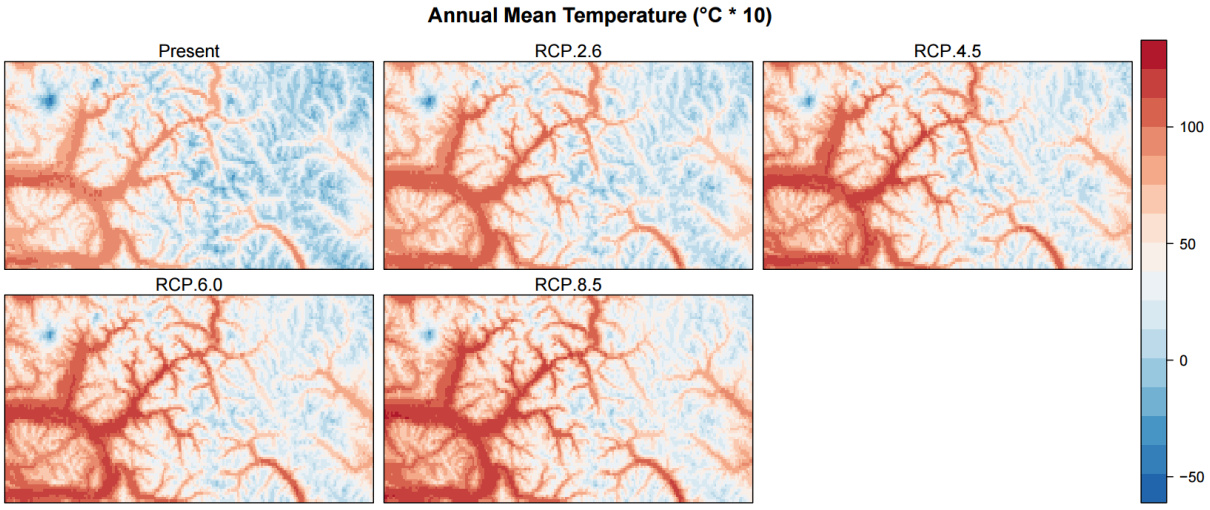

Figure 33: Washington annual mean temperature change

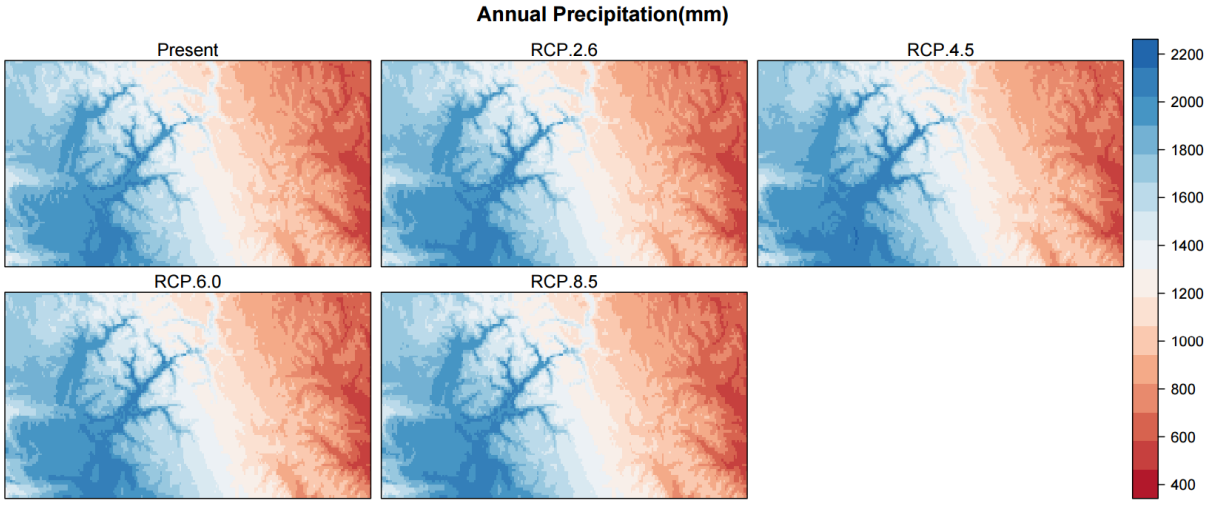

Figure 34: Washington annual precipitation change

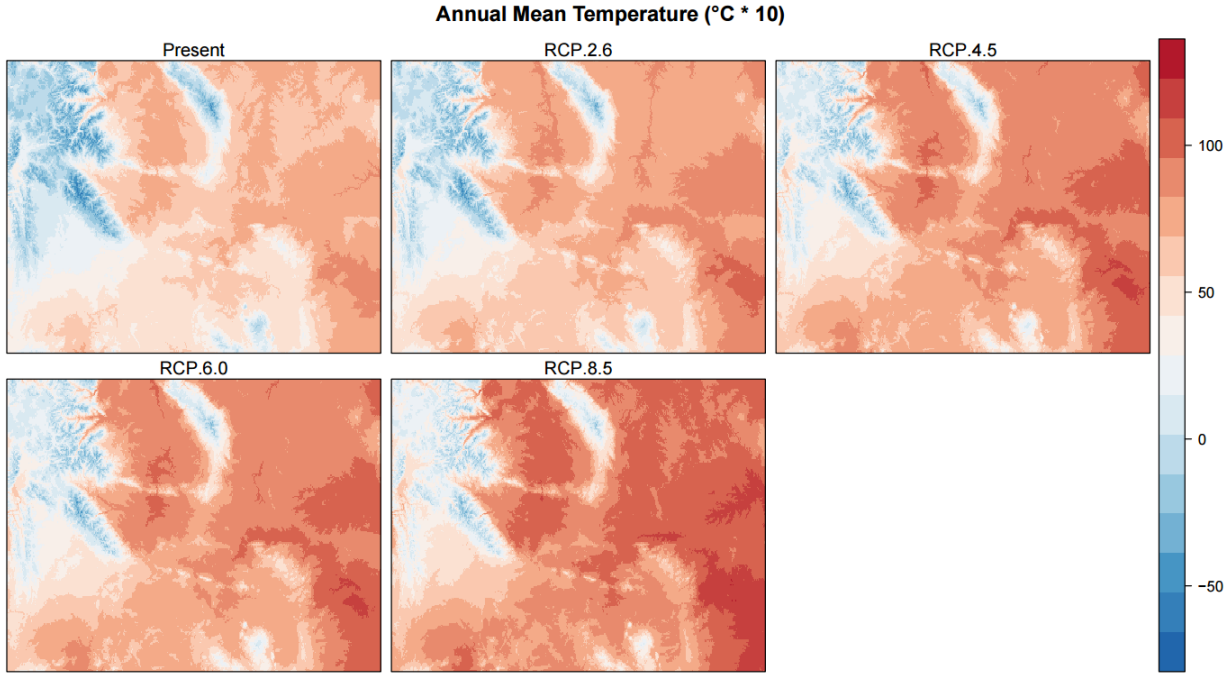

Figure 35: Wyoming annual mean temperature change

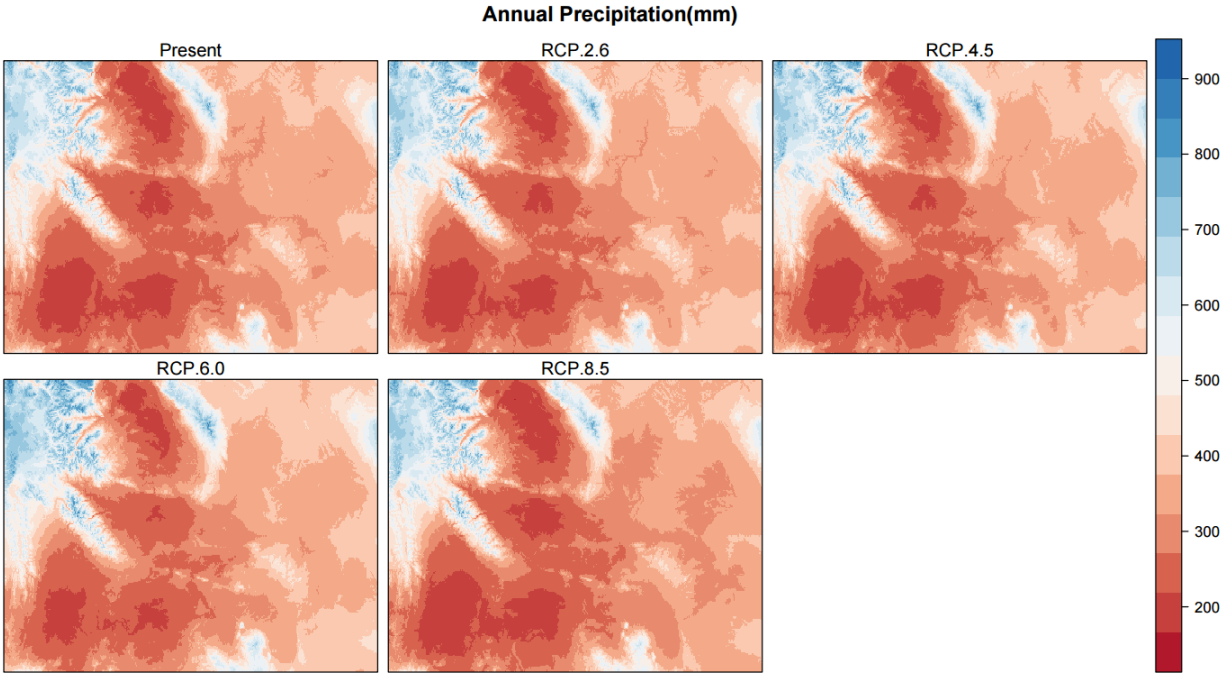

Figure 36: Wyoming annual precipitation change

6.2 Decision Tree and Land cover results

This section presents and describes the results obtained by using the C5.0 decision tree model to predict new land cover distribution maps for all four studied sites. Besides presenting the new maps, this section will also present the decision trees accuracy for correctly classifying land cover and what attributes are the best predictors for proper land cover classification.

6.2.1 New Mexico

The New Mexico training dataset is composed of 579,016 data cells. The subsequent trained decision tree scheme has a total of 26,464 leaf nodes, which are the possible outcomes of the tree.

Information regarding the importance of the attributes (predictors) to build the decision tree was obtained by individually measuring each attribute information gain ratio value, which is a method to measure relevancy of an attribute to properly split the target data, which in this case is the land cover. The most important attribute was elevation, with an information gain ratio value of 0.093. The second most important attribute was temperature with 0.085, precipitation was third with 0.077, slope was fourth with 0.059 and aspect proved to be the least important attribute, with a gain ratio of 0.005.

The final built tree was tested with a 10 fold cross-validation scheme, which resulted in 76% of the classes being correctly classified. Another test was performed, where the whole training data was fed into the tree without any folding scheme, resulting in an accuracy of 83%. Those tests proved that the tree is not overfitted for the training data, which deteriorates its capacity for correctly predicting new values. If the difference between the two tests methods was substantial, or if the cross-validation scheme returned a low accuracy value, it would mean that the decision tree is overfitted.

After validating the decision tree model, new land cover results were predicted by feeding the built tree with new future annual mean temperature and precipitation data. The five predicted results for land cover, four in the future and one in the present, the table and the graph with land cover change across all scenarios are displayed below.

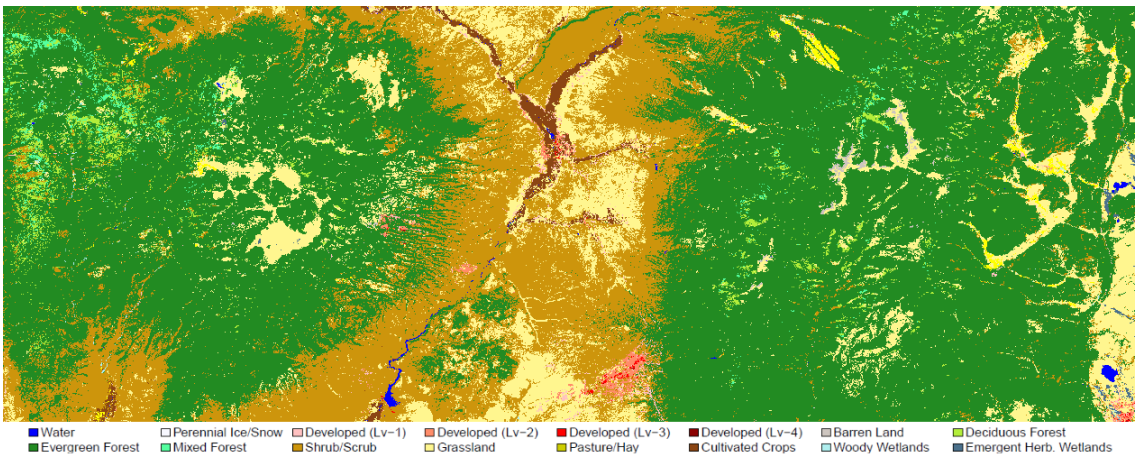

Figure 37: New Mexico - predicted land cover map for current climate

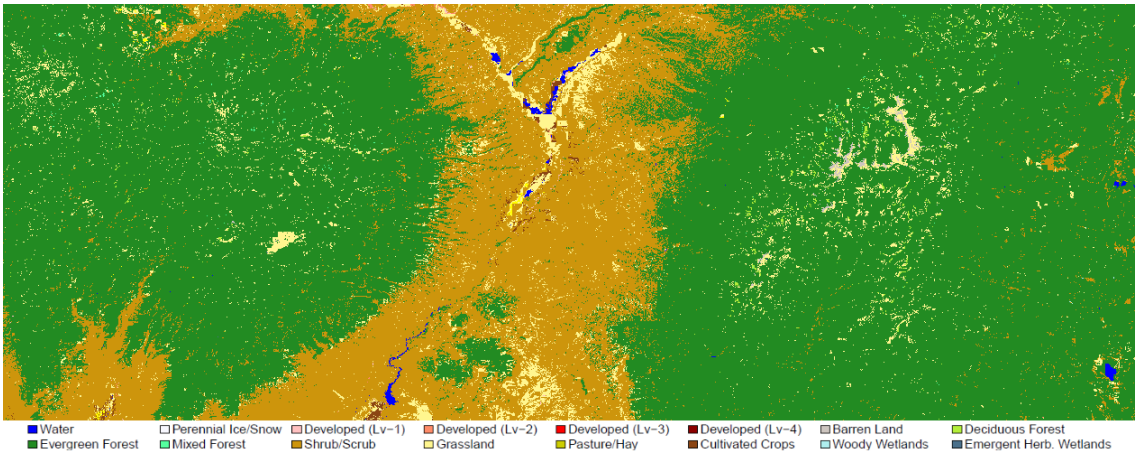

Figure 38: New Mexico - predicted land cover map for RCP 2.6 scenario

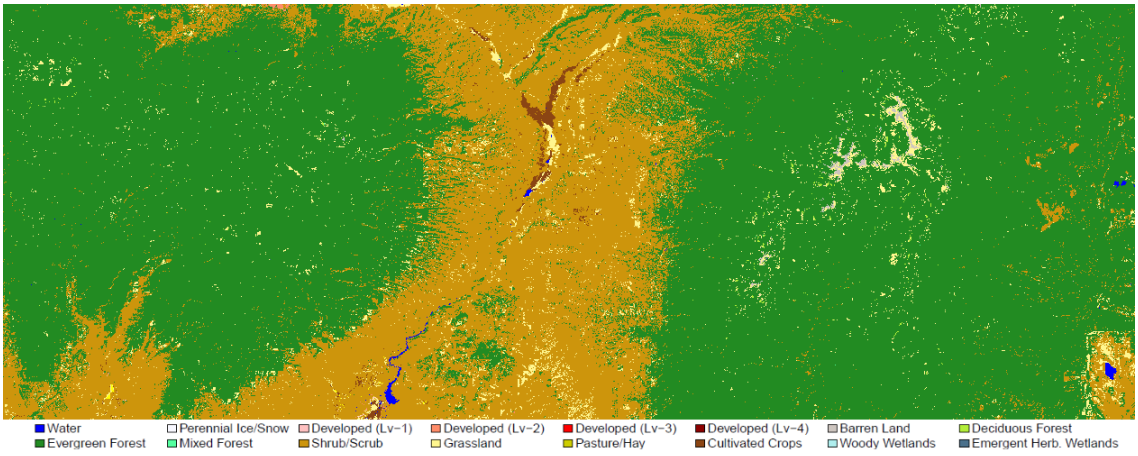

Figure 39: New Mexico - predicted land cover map for RCP 4.5 scenario

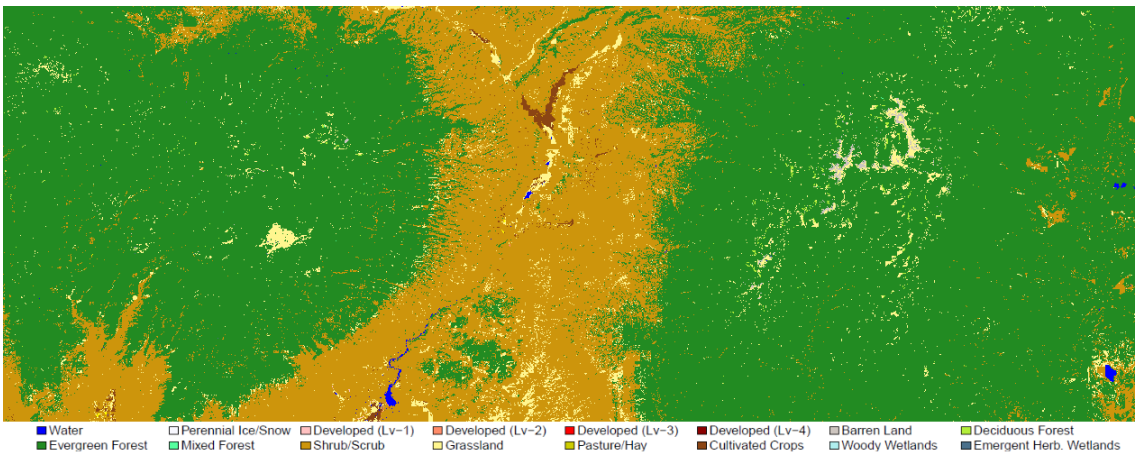

Figure 40: New Mexico - predicted land cover map for RCP 6.0 scenario

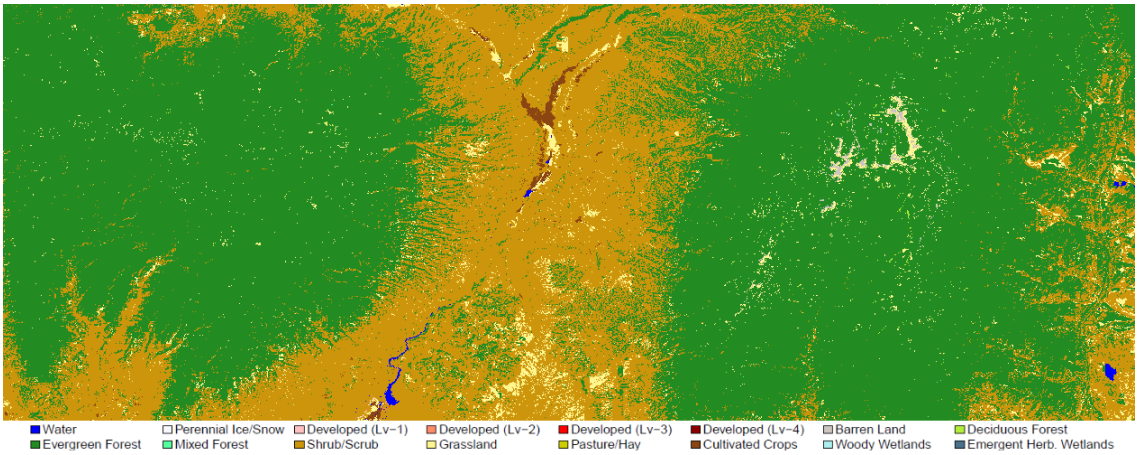

Figure 41: New Mexico - predicted land cover map for RCP 8.5 scenario

Table 4: New Mexico - Land cover class distribution

Classes	Present	RCP 2.6	RCP 4.5	RCP 6.0	RCP 8.5
Water	0.15%	0.19%	0.11%	0.11%	0.11%
Barren Land	0.17%	0.13%	0.16%	0.15%	0.16%
Deciduous Forest	1.58%	0.26%	0.10%	0.14%	0.08%
Evergreen Forest	62.56%	72.05%	71.30%	72.47%	68.19%
Mixed Forest	0.52%	0.04%	0.00%	0.01%	0.01%
Shrub/Scrub	21.92%	24.30%	26.93%	25.44%	30.00%
Grassland	12.25%	4.21%	2.47%	2.88%	2.60%
Cultivated Crops	0.85%	0.19%	0.36%	0.28%	0.34%

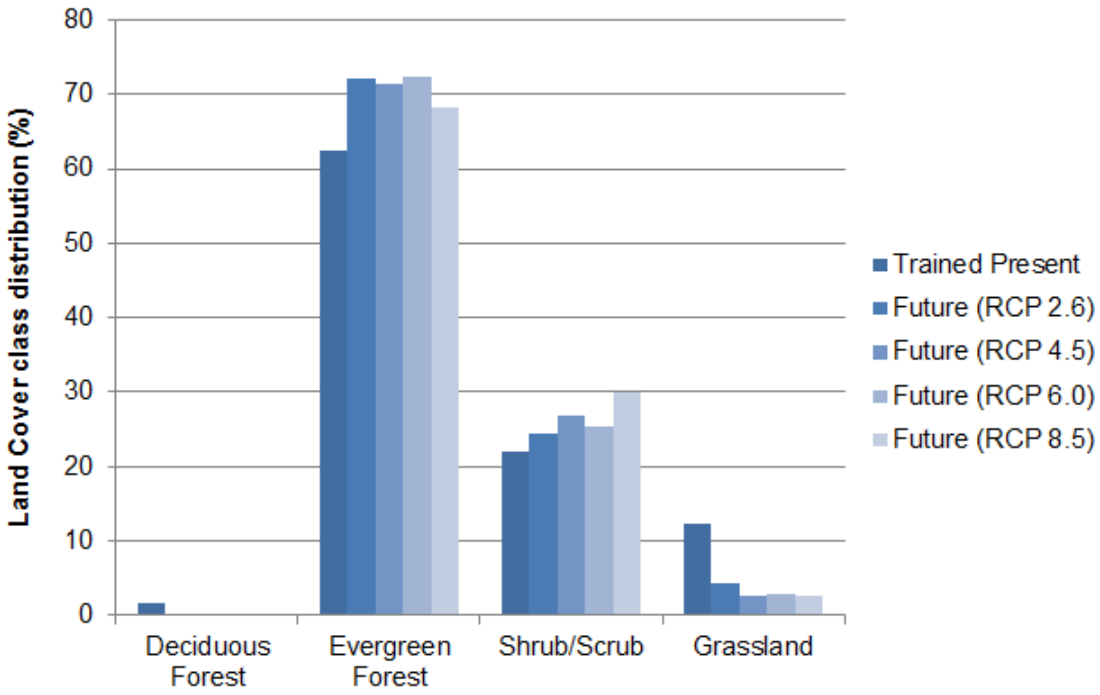

Figure 42: New Mexico - Land cover class distribution

The amount of land cover change between the predicted land cover for current climate and for the four future climate scenarios, RCP 2.6, RCP 4.5, RCP 6.0 and RCP 8.5 were 23.5%, 24.3%, 23.4% and 26.4%, respectively. The three most dominant vegetation classes in the region are evergreen forest, shrubs and grassland, respectively. Most of the land cover changes occur between those three classes.

In the RCP 2.6 scenario (*Figure* 38), the evergreen forest expanded over the grassland areas in the higher elevations of the Jemez Mountains (left mountain range located on the maps) and over the lower elevations on the right side of Santa Fe Mountains (right mountain range located on the maps). The shrubs covered most of the grasses in the lower elevation region between the two mountain ranges. Grassland cover decreases by 65.6%, while evergreen forest and shrubs increase by 15.2% and 10.9%, respectively.

In the RCP 4.5 scenario (*Figure* 39), most of the grassland areas located in the mountains have vanished because the evergreen forest expansion. The same occurs in the valley region at the center of the map, with the shrubs occupying more grassland areas. In this scenario, the areas covered by grassland are 79.8% lower than the current climate conditions, while evergreen forest and shrubs expanded by 14% and 22.9%, respectively, if compared to the land cover at current climate conditions.

Results for the RCP 6.0 scenario (*Figure* 40) are much like the RCP 4.5. Grassland areas are 76.5% lower than the current climate conditions and evergreen forest and shrubs areas are, respectively, 15.8% and 16.1% higher compared to the land cover at current climate conditions.

In the RCP 8.5 (*Figure* 41), the last tested climate scenario, the shrubs have expanded over a considerable area, covering most of the central area and the lower

elevations on the right side of Santa Fe Mountains. The area covered by shrubs is 36.9% greater than it was at current climate conditions. . Most of the grassland areas have vanished, accounting for a decrease of 78.8% of the original area. Evergreen forest cover is slightly greater than the original size, approximately 9%, although this is the scenario with the least amount of expansion.

6.2.2 Oregon

The Oregon site, consisting of 894,015 data cells in the input data, is the second largest dataset among the four studied locations. The resulting trained C5.0 decision tree scheme has a total of 15,509 leaf nodes. The Oregon tree size is the smallest among all sites because of the lack of different types of land cover, almost the whole area (97%) is covered by only two classes, which considerably reduces the amount of different outcomes for the tree.

Attribute importance ranking was obtained alongside the decision tree. The most important attribute was temperature, with an information gain ratio value of 0.064. The second most important attribute was elevation with 0.049, precipitation was third with 0.048, slope was fourth with 0.021 and aspect was the least important attribute, with a gain ratio of 0.003.

Testing the built tree with a 10 fold cross-validation scheme resulted in an accuracy of 80%, while testing with the whole training data, not dividing the data in multiple folds, resulted in an accuracy of 85%. These results show that the decision tree is not overfitted for the training data. Overall, the decision tree built for the Oregon site was the most accurate among all of the tested sites.

After testing the decision tree accuracy, new land cover results were predicted by feeding it new future annual mean temperature and precipitation data. The five predicted results for land cover, four in the future and one in the present, the table and the graph with land cover change across all scenarios are displayed below.

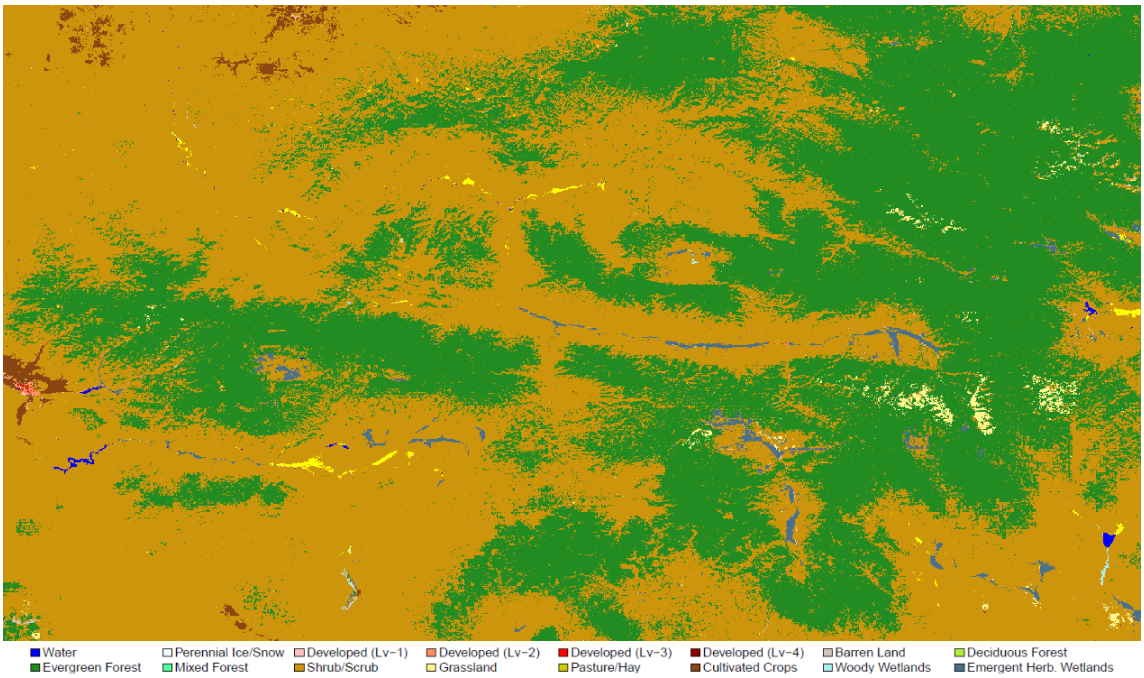

Figure 43: Oregon - predicted land cover map for current climate

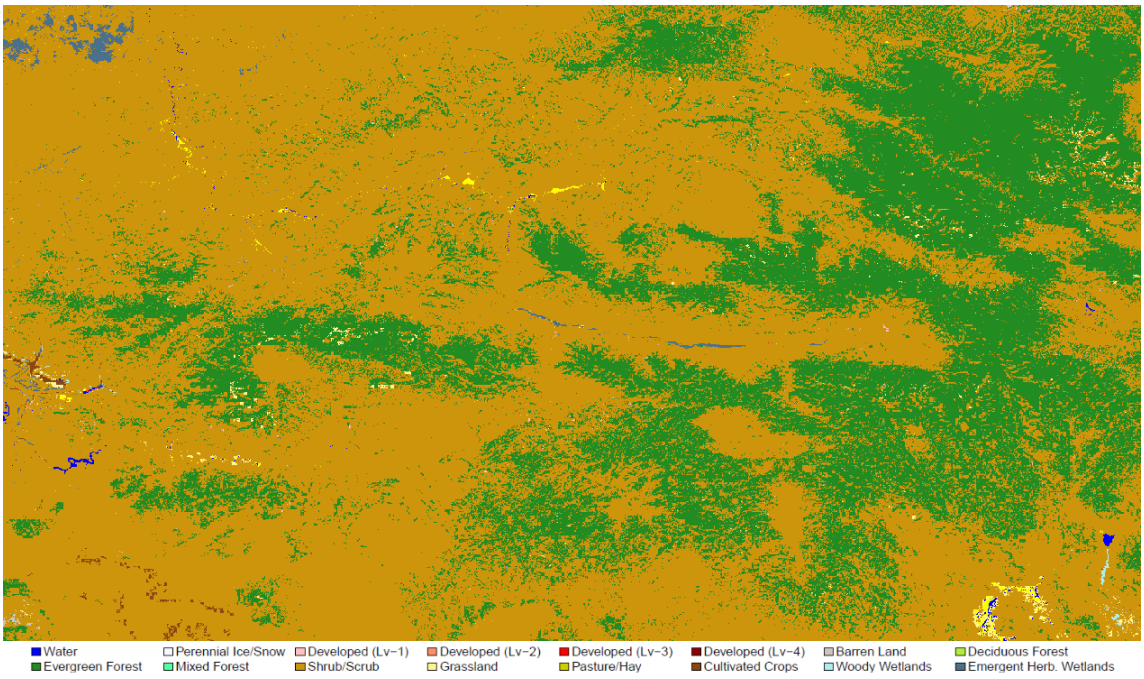

Figure 44: Oregon - predicted land cover map for RCP 2.6 scenario

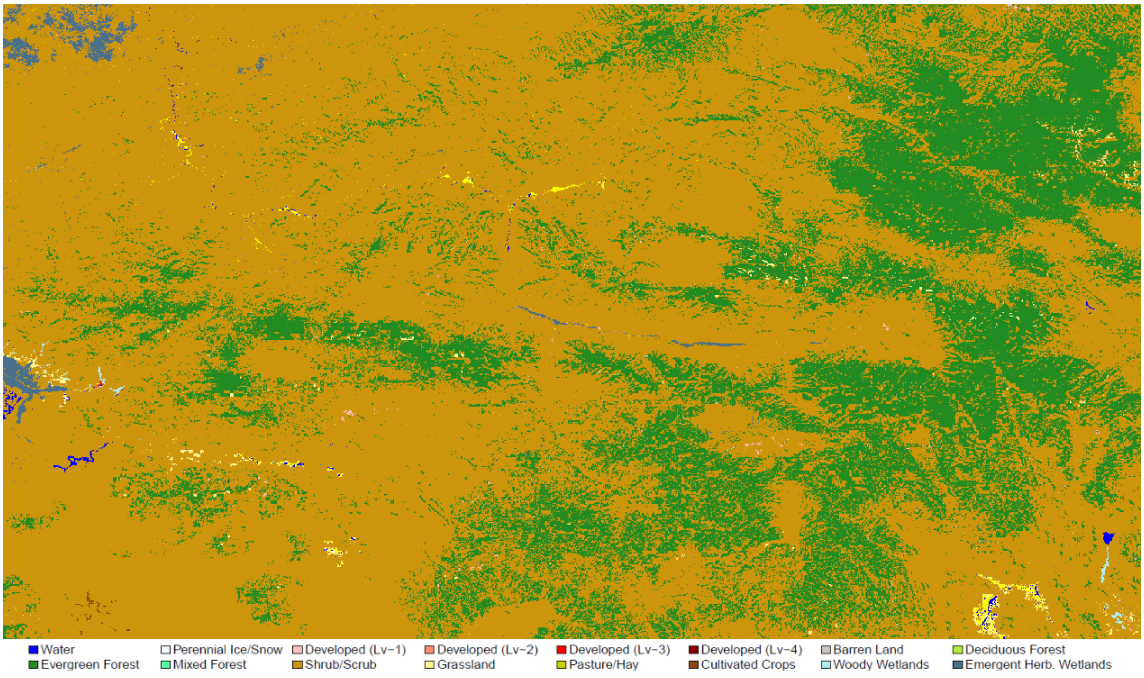

Figure 45: Oregon - predicted land cover map for RCP 4.5 scenario

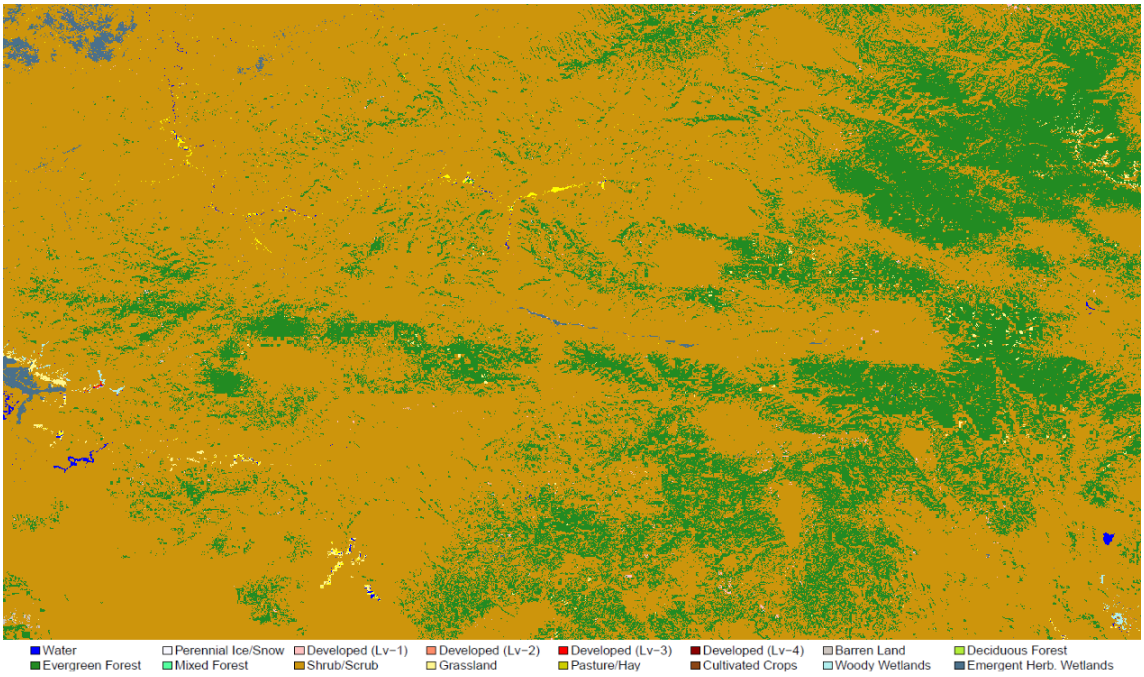

Figure 46: Oregon - predicted land cover map for RCP 6.0 scenario

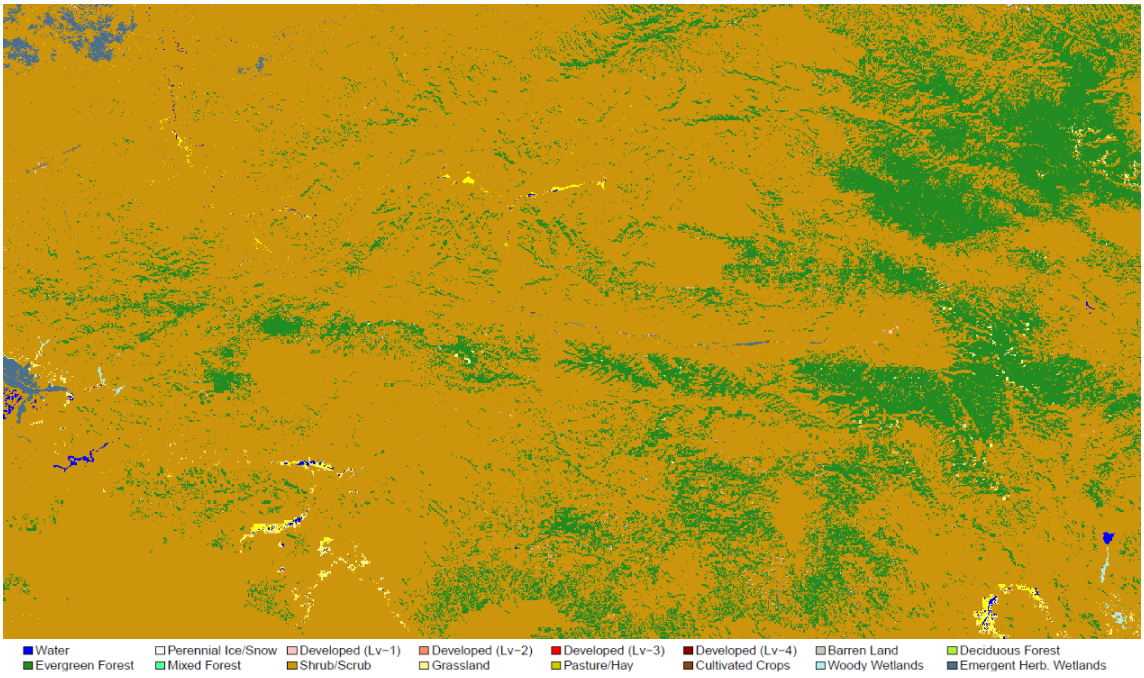

Figure 47: Oregon - predicted land cover map for RCP 8.5 scenario

Table 5: Oregon - Land cover class distribution

Classes	Present	RCP 2.6	RCP 4.5	RCP 6.0	RCP 8.5
Water	0.07%	0.10%	0.12%	0.09%	0.14%
Barren Land	0.03%	0.03%	0.01%	0.02%	0.01%
Evergreen Forest	39.88%	28.97%	23.82%	22.29%	17.57%
Shrub/Scrub	58.13%	69.88%	74.90%	76.58%	81.09%
Grassland	0.41%	0.31%	0.30%	0.24%	0.32%
Pasture/Hay	0.19%	0.16%	0.16%	0.10%	0.19%
Cultivated Crops	0.58%	0.13%	0.02%	0.00%	0.00%
Woody Wetlands	0.03%	0.03%	0.04%	0.04%	0.04%
Herbaceous Wetlands	0.68%	0.41%	0.62%	0.58%	0.60%

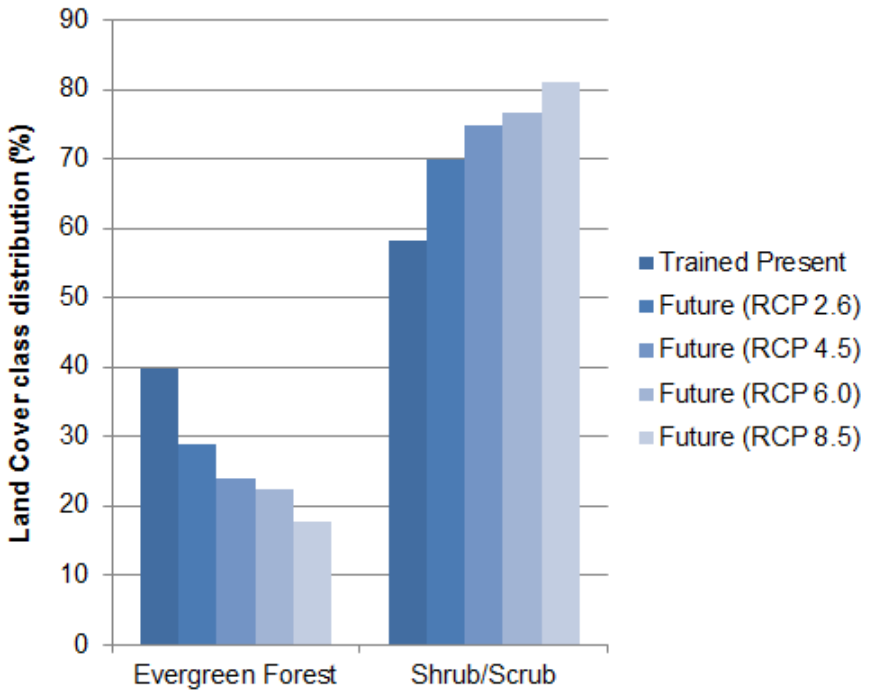

Figure 48: Oregon - Land cover class distribution

The amount of land cover change between the predicted land cover for current climate and for the four future climate scenarios, RCP 2.6, RCP 4.5, RCP 6.0 and RCP 8.5 were 21.7%, 25.8%, 26.8% and 29.8%, respectively.

Since over 97 % of the Oregon site is only represented by two land cover classes, evergreen forest and shrubs, the changes occurred almost exclusively between those classes. While the evergreen forest decreased, the shrubs expanded, superseding areas where there were evergreen forest. The rate of change of the evergreen forest across the four climate scenarios, RCP 2.6, RCP 4.5, RCP 6.0 and RCP 8.5, were -27.4%, -40.3%, -44.1%, and -56%, respectively. While the shrubs increased by 20.2%, 28.8%, 31.7%, and 39.5%, respectively for the same scenarios. The expansion of the shrub vegetation throughout the evergreen forest does not occur in a random manner, it follows the elevation gradient. Lower elevations were more suitable to change from forests to shrubs, while the evergreen forest located at the top of the mountains, mainly in the east and northeast of the region, was less affected by the warmer climate of the worst cases scenarios.

6.2.3 Washington

The Washington training dataset is composed of 626,934 data cells and the trained decision tree scheme has a total of 45,438 leaf nodes, being the second largest built tree among all of the four sites. The large tree is due to the fact that this region contains the highest amount of significant land cover classes, which considerable increased the quantity of possible outcomes for the classification tree.

Attribute importance ranking was obtained alongside the decision tree. The most important attribute was elevation, with an information gain ratio value of 0.084. The

second most important attribute was temperature with 0.067, slope was third with 0.039, precipitation was fourth with 0.033 and aspect was the least important, with a gain ratio of 0.023.

The final built tree was tested with a 10 fold cross-validation scheme, which resulted in 70% of the classes being correctly classified. A second test was performed, where the entire training data was fed into tree without any folding scheme, resulting in an accuracy of 82%. The cross-validation accuracy did not achieve a higher value, like the ones achieved on the other sites, because of the large size of the tree; however, 70% accuracy is acceptable.

After testing its accuracy, the built classification tree is used to predict new land cover results by feeding it with new annual mean temperature and precipitation data. The predicted land cover maps, the table and graph with the land cover class distribution across all scenarios are displayed below.

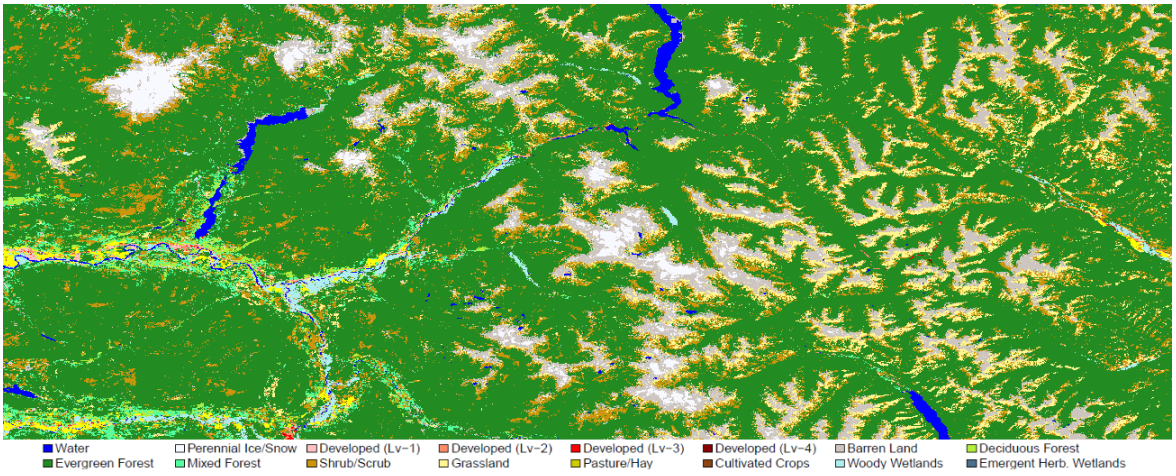

Figure 49: Washington - predicted land cover map for current climate

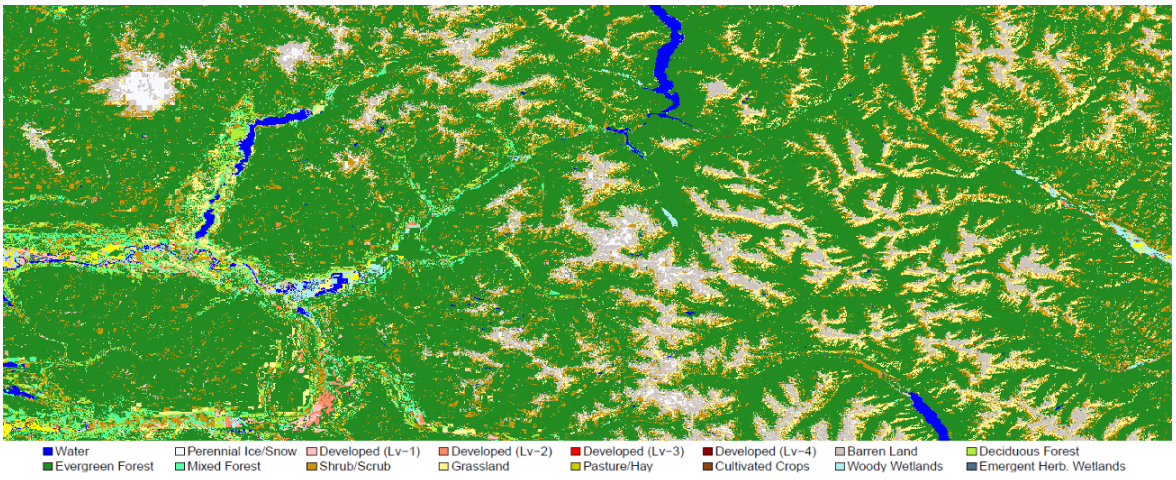

Figure 50: Washington - predicted land cover map for RCP 2.6 scenario

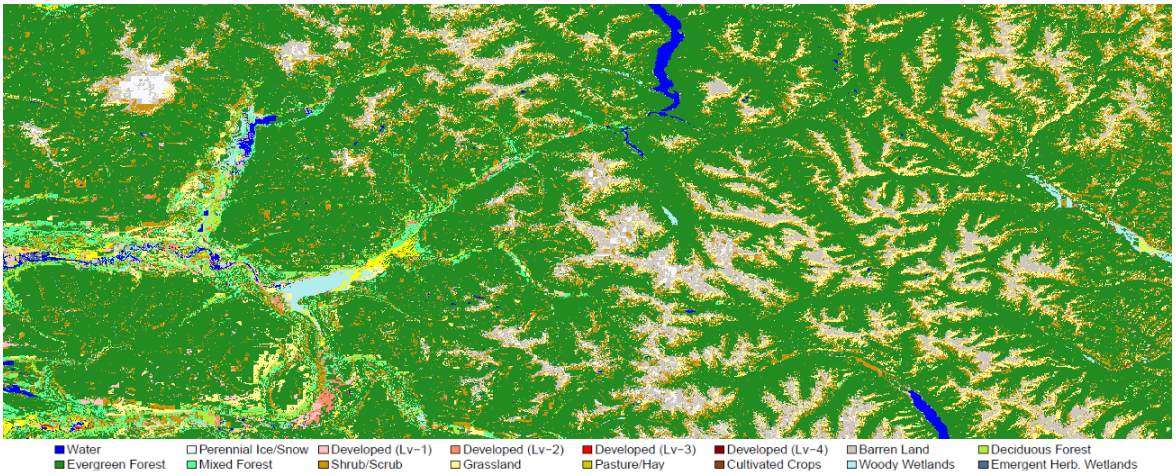

Figure 51: Washington - predicted land cover map for RCP 4.5 scenario

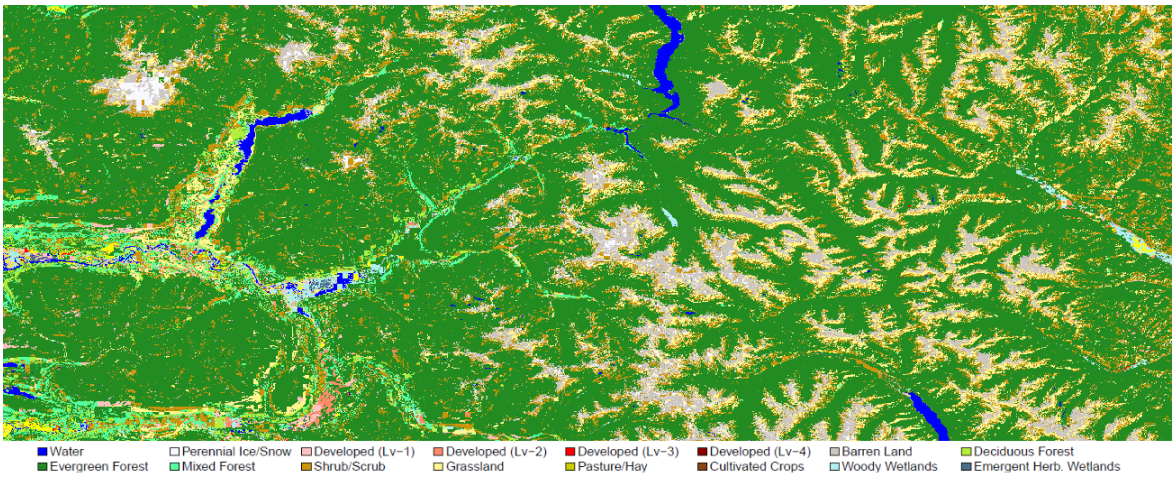

Figure 52: Washington - predicted land cover map for RCP 6.0 scenario

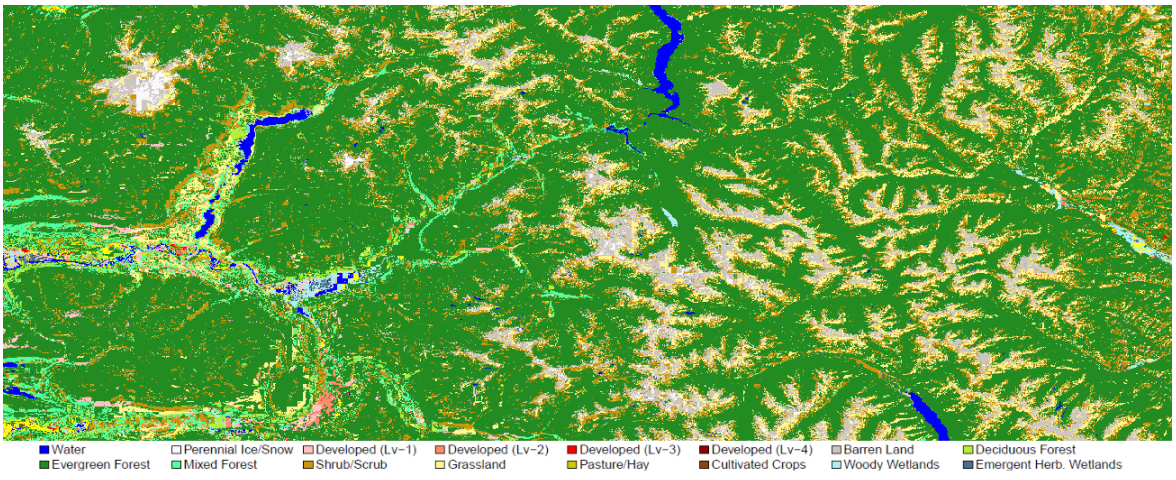

Figure 53: Washington - predicted land cover map for RCP 8.5 scenario

Table 6: Washington - Land cover class distribution

Classes	Present	RCP 2.6	RCP 4.5	RCP 6.0	RCP 8.5
Water	0.99%	0.97%	0.78%	0.92%	0.94%
Perennial Ice/Snow	1.76%	1.06%	0.90%	0.78%	0.63%
Barren Land	7.70%	6.87%	6.66%	6.35%	6.02%
Deciduous Forest	0.87%	1.34%	1.37%	1.31%	1.26%
Evergreen Forest	68.80%	68.49%	67.06%	67.42%	66.70%
Mixed Forest	1.42%	1.68%	1.92%	1.97%	2.12%
Shrub/Scrub	10.41%	10.66%	11.01%	11.26%	11.85%
Grassland	6.71%	7.70%	8.68%	8.66%	9.09%
Pasture/Hay	0.45%	0.30%	0.38%	0.29%	0.28%
Cultivated Crops	0.01%	0.01%	0.01%	0.00%	0.01%
Woody Wetlands	0.83%	0.58%	0.73%	0.60%	0.58%
Herbaceous Wetlands	0.04%	0.05%	0.04%	0.06%	0.07%

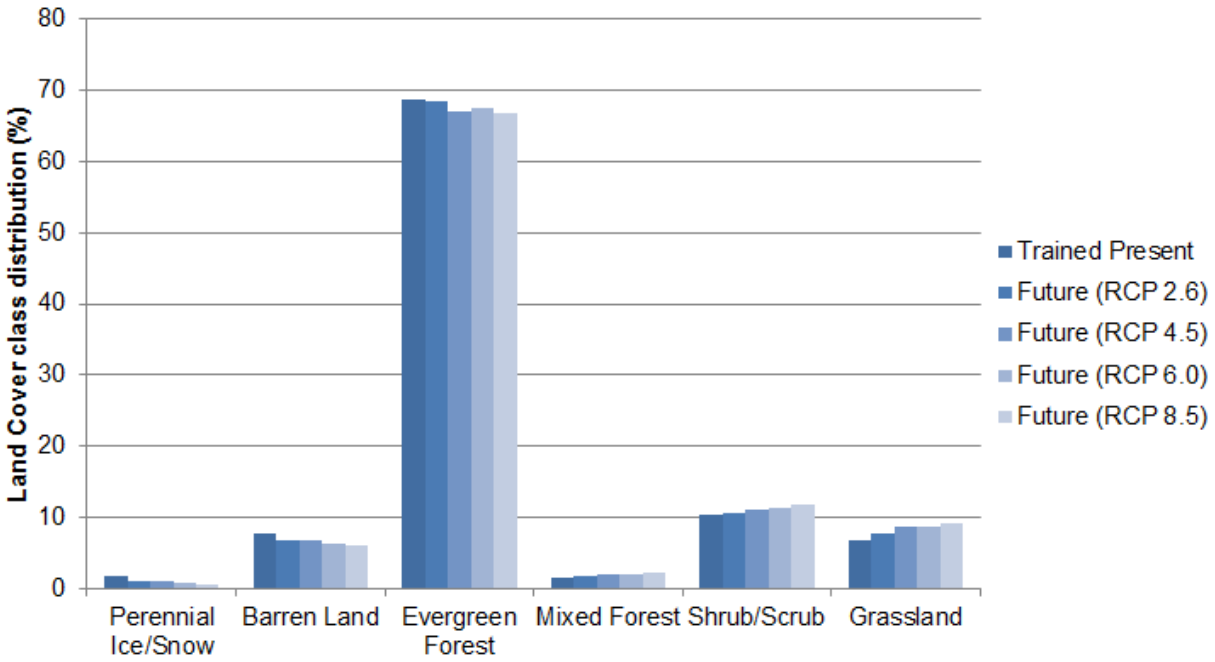

Figure 54: Washington - Land cover class distribution

The Washington region is the site that presents the highest amount of significant distinguished land cover classes. Perennial ice and snow, barren land (rock outcrop formations), evergreen forest, shrubs and grassland land cover classes account for more than 95% of the total area of the region. The amount of land cover change between the predicted land cover for current climate and for the four future climate scenarios, RCP 2.6, RCP 4.5, RCP 6.0 and RCP 8.5 were 25.8%, 30.0%, 29.1% and 29.4%, respectively.

The perennial ice and snow cover, present exclusively on the peaks of the Northern Cascade Mountains, is the most sensitive land cover class to change with a warming environment. Results show that the perennial snow cover decreases by 39.8%, 48.9%, 55.5% and 64%, respectively, for the four climate scenarios; RCP 2.6, RCP 4.5, RCP 6.0 and RCP 8.5. It can be observed in the future land cover maps that the snow present in the mountains located in the central region has almost completely vanished and most of the snow and ice present at Mount Baker (northwest part of the map) has disappeared.

The barren land areas, composed of rock outcrops, are located at the top of the mountains as well as beneath and around the snow covered areas. Those areas were also affected; the decrease was 10.7%, 13.4%, 17.5% and 21.8%, respectively, for the four climate scenarios. Although the decrease on snow cover exposed more the barren land, expanding its area, the expansion of the surrounding evergreen forest, grassland and shrubs over the rock formations were more significant.

Evergreen forest cover area has slightly decreased by a rate of 0.4%, 2.5%, 2% and 3.1%, respectively, for the four climate scenarios. While these changes are small in value, the evergreen forest accounts for more than 60% of the territory. The majority of those areas were covered by the shrubs and the grasslands. The grasslands have

increased by 14.7%, 29.2%, 28.9%, and 35.5%, while the shrubs increased by 2.4%, 5.7%, 8.2%, and 13.9% respectively, for the four climate scenarios.

6.2.4 Wyoming

The Wyoming state site, consisting of 963,858 data cells in the input data, is the largest dataset among the four studied ones and, as a consequence of its vast amount of data, it also presents the largest trained decision tree scheme with 54,125 leaf nodes (possible outcomes). In order to assess the accuracy of the model, the built tree was tested with a 10 fold cross-validation scheme, which resulted in an accuracy of 69%. An accuracy of 80% was achieved after testing the model over the whole training data. The same justification of the Washington site can be applied to this case. The substantial size of the tree undermines its accuracy.

Once again, the information gain ratio values were extracted to determine the importance of the predictors. Temperature was the most important attribute, with an information gain ratio value of 0.066; elevation was second (0.060), precipitation third (0.058), slope was fourth (0.050) and aspect last (0.006).

The predicted land cover maps, the table and graph with the land cover class distribution across all scenarios are displayed below.

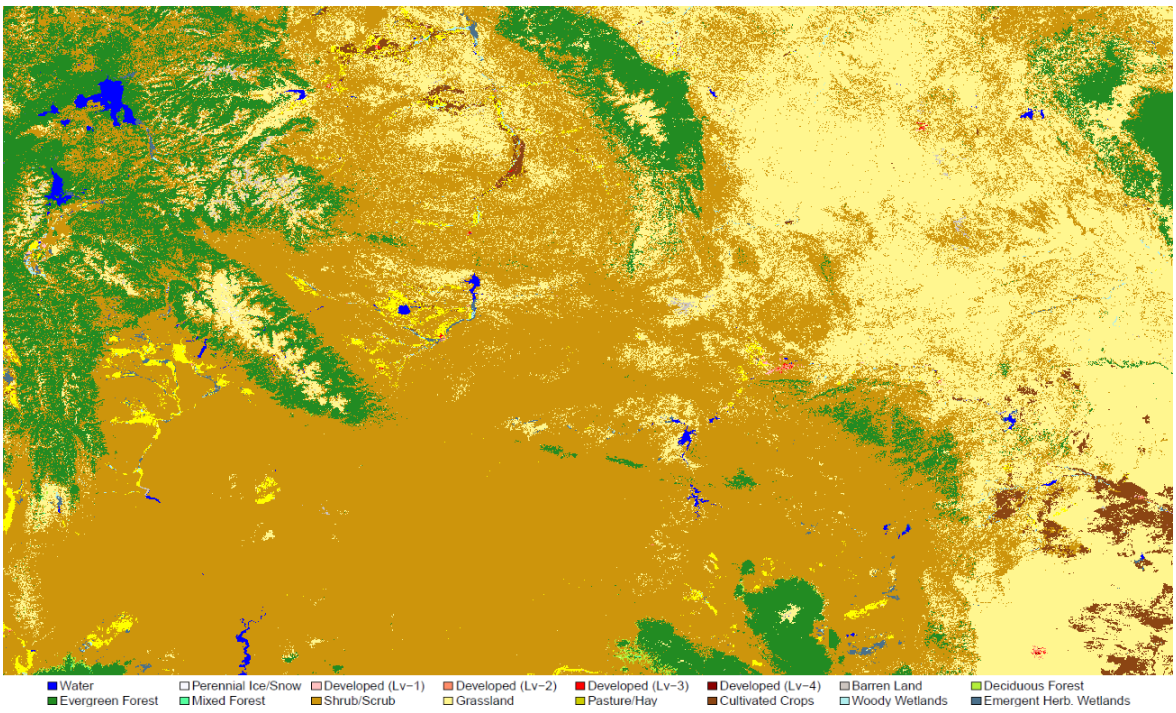

Figure 55: Wyoming - predicted land cover map for current climate

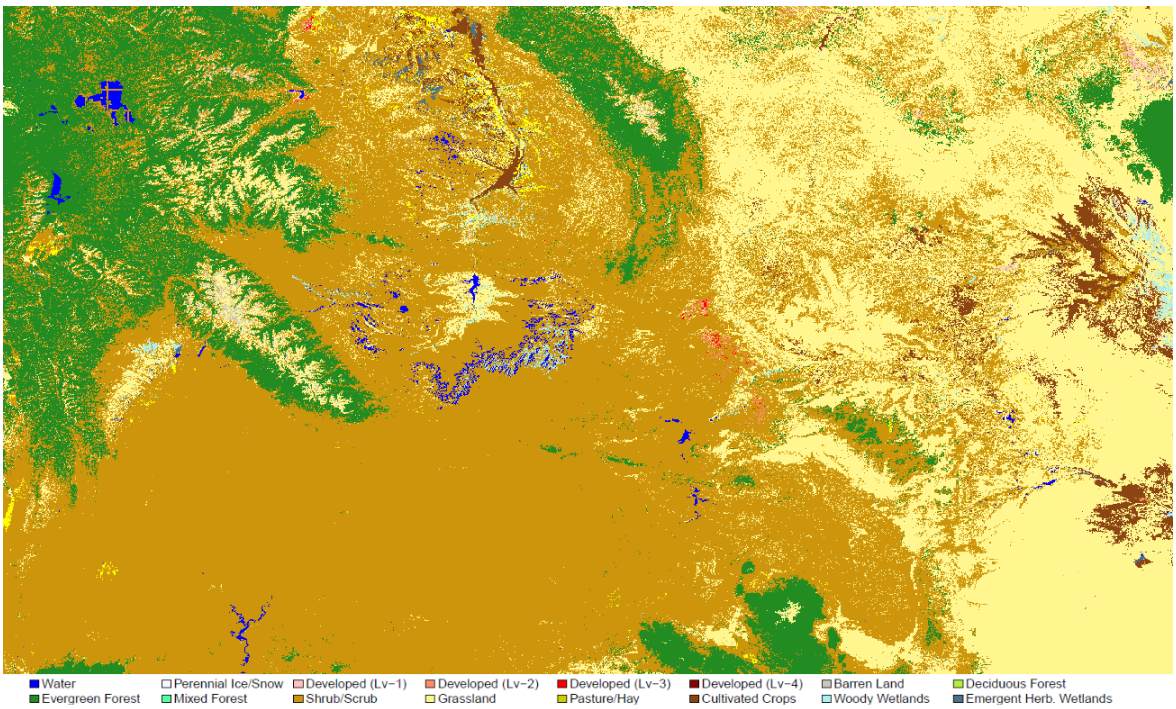

Figure 56: Wyoming - predicted land cover map for RCP 2.6 scenario

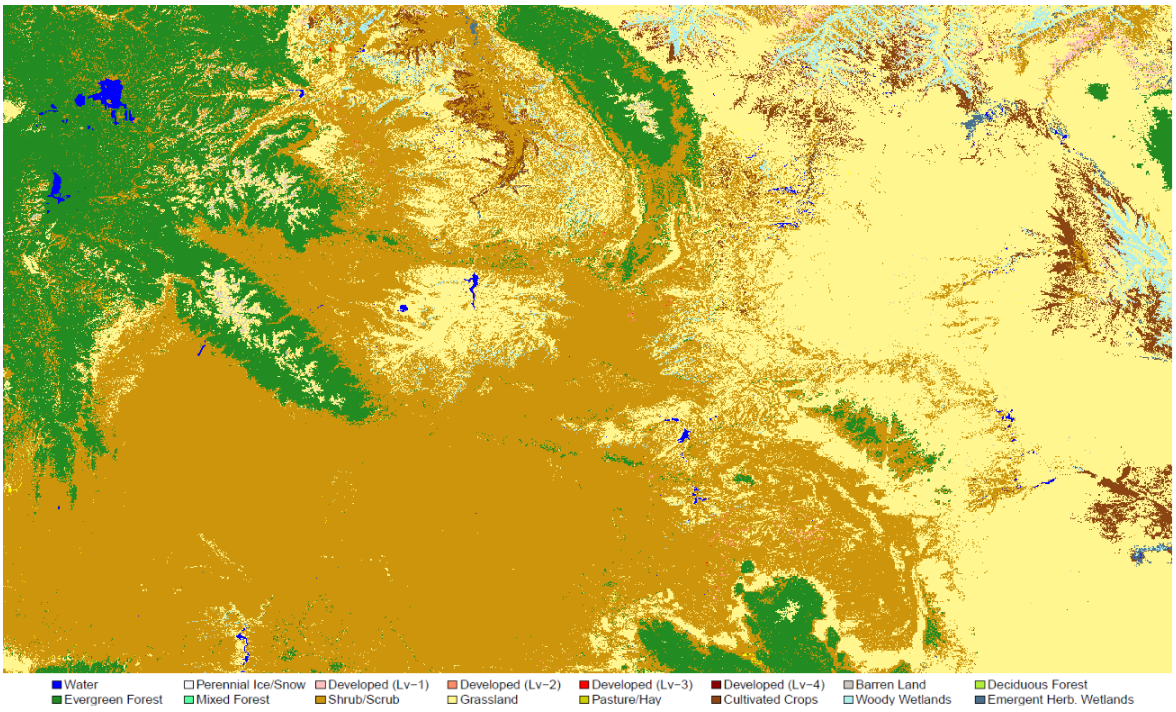

Figure 57: Wyoming - predicted land cover map for RCP 4.5 scenario

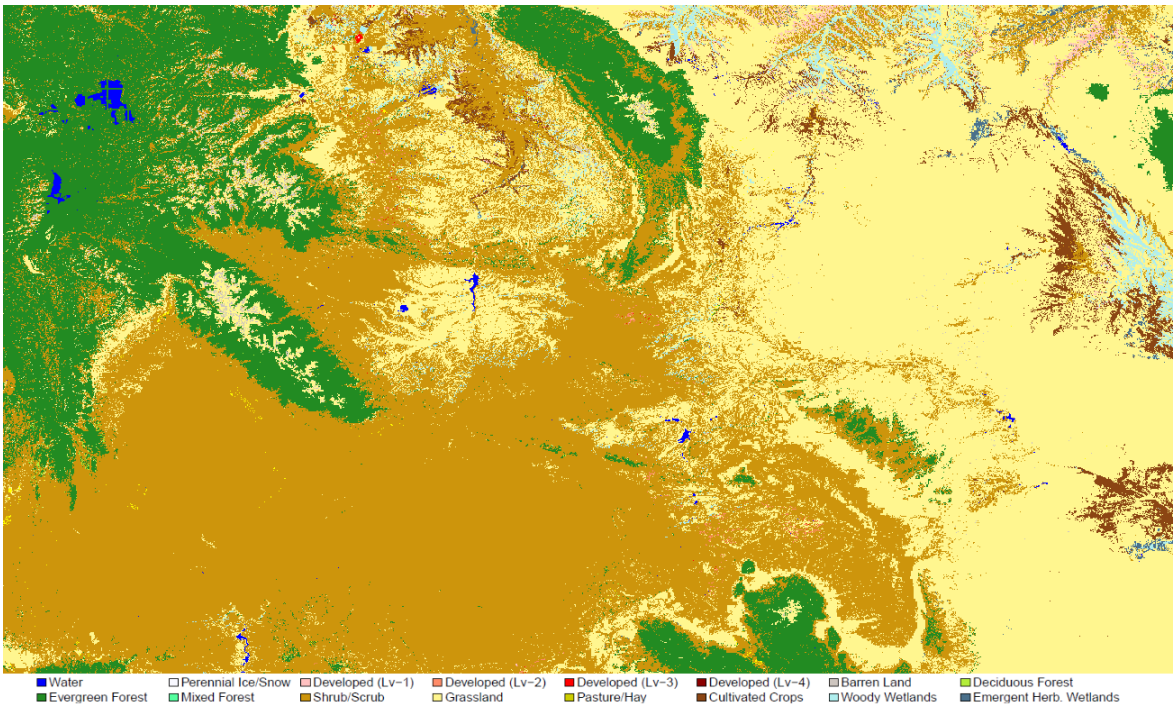

Figure 58: Wyoming - predicted land cover map for RCP 6.0 scenario

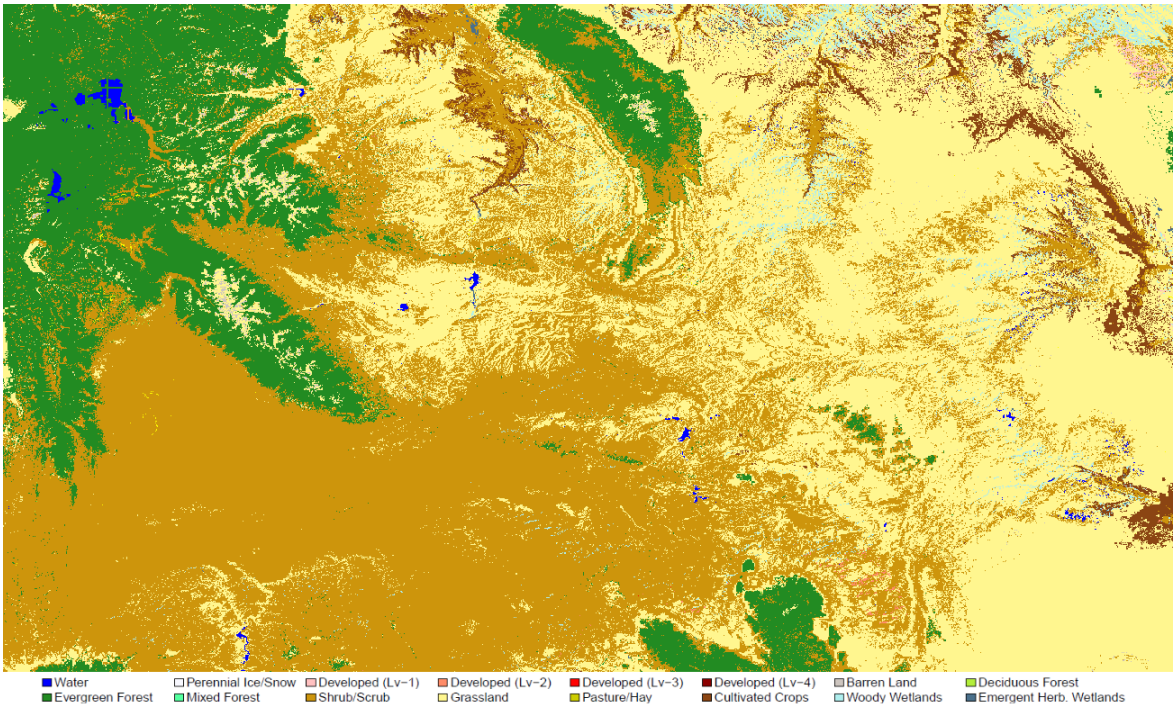

Figure 59: Wyoming - predicted land cover map for RCP 8.5 scenario

Table 7: Wyoming - Land cover class distribution

Classes	Present	RCP 2.6	RCP 4.5	RCP 6.0	RCP 8.5
Water	0.40%	0.56%	0.29%	0.25%	0.26%
Barren Land	0.27%	0.25%	0.28%	0.31%	0.15%
Deciduous Forest	0.11%	0.03%	0.06%	0.06%	0.03%
Evergreen Forest	13.16%	14.29%	15.14%	15.76%	15.03%
Shrub/Scrub	55.28%	52.62%	43.63%	42.36%	40.97%
Grassland	28.40%	29.25%	36.46%	37.47%	39.61%
Pasture/Hay	1.00%	0.40%	0.07%	0.09%	0.05%
Cultivated Crops	1.21%	2.07%	2.27%	1.85%	2.53%
Woody Wetlands	0.18%	0.49%	1.49%	1.40%	1.35%

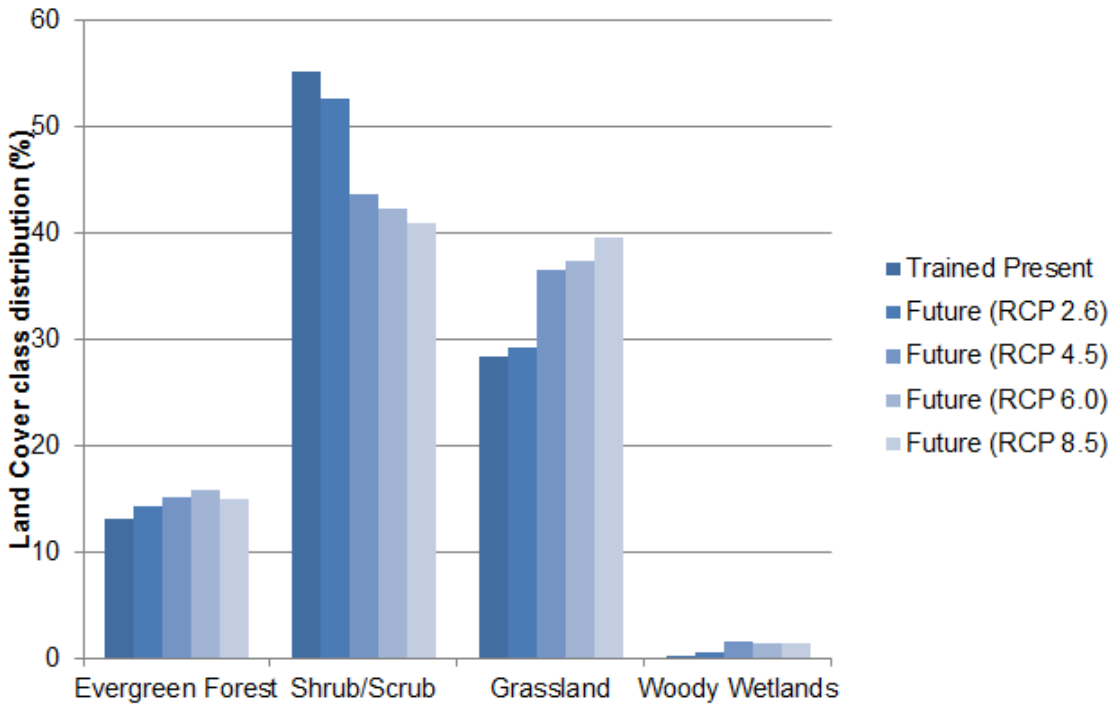

Figure 60: Wyoming - Land cover class distribution

Wyoming was the site that presented the most extensive land cover change between the predicted land cover for current climate and the four future climate scenarios, RCP 2.6, RCP 4.5, RCP 6.0 and RCP 8.5. The land cover modification accounted for was 30.4%, 33.3%, 33.5% and 36.8%, respectively.

The most notable difference between the current and future land cover is the expansion of the short mixed-grass prairie (grassland) of the Great Plains from the east part of Wyoming towards the center of the state. This expansion is more evident in the RCP 4.5, RCP 6.0 and RCP 8.5 scenarios. Grassland areas increased by 3%, 28.4%, 31.9% and 39.5%, respectively for the four climate scenarios. As the grassland areas expands, the shrub regions decreases. Decrease rates are 4.8%, 21.1%, 23.4% and 25.9%, respectively for the four scenarios.

Evergreen forest also expanded along the mountain ranges over areas covered by grasslands and shrubs, mainly in the higher altitudes of Wind River and Teton mountain ranges, located in the northwest of the state. Evergreen forests increased by 8.6%, 15%, 19.8% and 14.3%, respectively for the four scenarios.

CHAPTER VII

CONCLUSION

Results obtained in this study revealed that the relation between biogeography and climate is relatively strong. Major alterations in the climate can deeply affect terrestrial vegetation and plant species distribution. All studied sites presented significant natural environmental modifications. It is important to state that the scope of this work is to predict only natural modifications caused by climate, therefore modifications caused by human actions, like deforestation and land use change cannot be accounted for.

Each site presented different types of modifications for future terrestrial ecosystems. For New Mexico, predictions showed that the shrubs will expand over grassland areas as well as part of the evergreen forest. The results obtained for the Oregon site demonstrated a significant expansion of shrubs over evergreen forest. For the Washington site, the significant loss of perennial snow is quite remarkable and alarming, shrubs and grasslands also expanded over areas dominated by evergreen forest. Lastly, the results obtained in Wyoming revealed that the short mixed-grass prairie will expand considerably over shrub land areas.

The use of decision trees to predict the modification between future and current natural land cover has been successfully applied in this study. Decision trees have proven to be a fast and reliable alternative to other highly computational simulation methods, since it presents the ability to handle enormous amount of data cells; which is a limitation for most of the traditional vegetation simulation models (Cramer et al., 2001; Bachelet, 2001).

The model performed well for all four locations, achieving prediction accuracies for current land cover of 83%, 85%, 82% and 80% respectively for New Mexico, Oregon, Washington and Wyoming sites. The fact that the sites present great differences in size but managed to attain high classification accuracy corroborate the versatility of the algorithm to handle efficient modeling, regardless of the size of the region. So, the model could be perfectly suitable for applications involving regional and global scales, if the proper resolution of the dataset is chosen.

Another advantage of the decision tree model is its flexibility in using different types of data. Although only a few datasets were used in this work, many others could be used for training the model. The accuracy achieved using only five datasets, specifically elevation, aspect, slope, temperature and precipitation, was remarkable; proving that the model can perform well even if only a few datasets are available to be used, as long as they are statistically significant to the target class, which in our case, is land cover.

One limitation of the model is the assumption that the biogeographic rules governing the relations between climate, topography and vegetation and species distribution, will remain the same or relatively close in the future. The decision tree model can only learn the rules for present climate conditions; the prediction for future vegetation distribution is just the application of those same rules for a different, warmer environment. So, if vegetation rapidly adapts to this new environment, and remain the same type without changing its type for a specific location, i.e. forest to shrubs, the prediction based on the decision tree model might not be accurate. The results obtained in Oregon will be used to better illustrate this statement. The evergreen forest in the Oregon site has drastically decreased across the four climate scenarios, while the shrubs have expanded and taken territories that belonged to the forest. This occurred because shrub vegetation is highly correlated to a warmer climate and

the forest to a cooler one. However, if this correlation does not sustain as same in the future climate, the prospective land cover might not be altered as much as was predicted.

However, previous researches using models that relate vegetation distribution and climate, like the DVGM, to predict future natural land cover driven by temperature and precipitation change have resulted in significant land cover modification (Hayhoe et al., 2004; Cramer et al., 2001; Huntley et al., 1995; Hickler et al., 2012), which validates the theory that terrestrial ecosystems are highly sensitive to climate change. Therefore, using machine learning techniques, such as decision trees, can be a reliable alternative for developing models that predict vegetation distribution in the future. This thesis, as well as many other studies (McIver and Friedl, 2002; Sesnie et al., 2008; de Colstoun et al., 2003; Friedl and Brodley, 1997; Hansen et al., 2000), have proven that decision tree models can be powerful techniques to predict current land cover and future land cover change. There are many other new machine learning models that can also be applied to remote sensing problems that need further studies.

Although this thesis only studied the prediction of future land cover based on mean annual climate data, for both current and future conditions, the model could also be applied to a more seasonal scope, such as prediction of the snow cover change over future winter seasons. It could also be used to predict future patterns of snow formation and melting.

In conclusion, the model has demonstrated to be a computation efficient method that can be applied over large areas and preserve the high resolution attribute information.

APPENDIX A

NATIONAL LAND COVER DATABASE 2001 (NLCD2001) LEGEND


Open Water: All areas of open water, generally with less than 25 percent cover of vegetation or soil.

Perennial Ice/Snow: All areas characterized by a perennial cover of ice and/or snow, generally greater than 25 percent of total cover.

Developed, Open Space: Includes areas with a mixture of some constructed materials, but mostly vegetation in the form of lawn grasses. Impervious surfaces account for less than 20 percent of total cover. These areas most commonly include large-lot single-family housing units, parks, golf courses, and vegetation planted in developed settings for recreation, erosion control, or aesthetic purposes.


Developed, Low Intensity: Includes areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 20-49 percent of total cover. These areas most commonly include single-family housing units.


Developed, Medium Intensity: Includes areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 50-79 percent of the total cover. These areas most commonly include single-family housing units.


Developed, High Intensity: Includes highly developed areas where people reside or work in high numbers. Examples include apartment complexes, row houses, and

commercial/industrial. Impervious surfaces account for 80 to 100 percent of the total cover.

 Barren Land (Rock/Sand/Clay): Barren areas of bedrock, desert pavement, scarps, talus, slides, volcanic material, glacial debris, sand dunes, strip mines, gravel pits, and other accumulations of earthen material. Generally, vegetation accounts for less than 15 percent of total cover.

 Deciduous Forest: Areas dominated by trees generally greater than 5 meters tall, and greater than 20 percent of total vegetation cover. More than 75 percent of the tree species shed foliage simultaneously in response to seasonal change.

 Evergreen Forest: Areas dominated by trees generally greater than 5 meters tall, and greater than 20 percent of total vegetation cover. More than 75 percent of the tree species maintain their leaves all year. Canopy is never without green foliage.

 Mixed Forest: Areas dominated by trees generally greater than 5 meters tall, and greater than 20 percent of total vegetation cover. Neither deciduous nor evergreen species are greater than 75 percent of total tree cover.

 Shrub/Scrub: Areas dominated by shrubs; less than 5 meters tall with shrub canopy typically greater than 20 percent of total vegetation. This class includes true shrubs, young trees in an early successional stage, or trees stunted from environmental conditions.

 Grassland/Herbaceous: Areas dominated by graminoids or herbaceous vegetation, generally greater than 80 percent of total vegetation. These areas are not subject to intensive management such as tilling, but can be utilized for grazing.

 Pasture/Hay: Areas of grasses, legumes, or grass-legume mixtures planted for livestock grazing or the production of seed or hay crops, typically on a perennial cycle. Pasture/hay vegetation accounts for greater than 20 percent of total vegetation.

 Cultivated Crops: Areas used for the production of annual crops, such as corn, soybeans, vegetables, tobacco, and cotton, and also perennial woody crops such as orchards and vineyards. Crop vegetation accounts for greater than 20 percent of total vegetation. This class also includes all land being actively tilled.

 Woody Wetlands: Areas where forest or shrubland vegetation accounts for greater than 20 percent of vegetative cover and the soil or substrate is periodically saturated with or covered with water.

 Emergent Herbaceous Wetlands: Areas where perennial herbaceous vegetation accounts for greater than 80 percent of vegetative cover and the soil or substrate is periodically saturated with or covered with water.

Bibliography

A. Anav, F. D'Andrea, N. Viovy, and N. Vuichard. A validation of heat and carbon fluxes from high-resolution land surface and regional models. *Journal of Geophysical Research: Biogeosciences (2005–2012)*, 115(G4), 2010.

M. Austin. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological modelling*, 157(2):101–118, 2002.

D. Bachelet. *MC1: A dynamic vegetation model for estimating the distribution of vegetation and associated ecosystem fluxes of carbon, nutrients, and water*. DIANE Publishing, 2001.

C. Beer, W. Lucht, D. Gerten, K. Thonicke, and C. Schmullius. Effects of soil freezing and thawing on vegetation carbon density in siberia: A modeling analysis with the lund-potsdam-jena dynamic global vegetation model (lpj-dgvm). *Global Biogeochemical Cycles*, 21(1), 2007.

C. M. Bishop et al. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

G. B. Bonan. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *science*, 320(5882):1444–1449, 2008.

V. Brovkin, S. Sitch, W. Von Bloh, M. Claussen, E. Bauer, and W. Cramer. Role of land cover changes for atmospheric co2 increase and climate change during the last 150 years. *Global Change Biology*, 10(8):1253–1266, 2004.

P. A. Burrough and R. A. McDonnell. *Principles of geographical information Systems*, volume 19988. Oxford University Press, 2011.

J. G. Carbonell, R. S. Michalski, and T. M. Mitchell. An overview of machine learning. In *Machine learning*, pages 3–23. Springer, 1983.

T. Carter, M. Hulme, and M. Lal. General guidelines on the use of scenario data for climate impact and adaptation assessment. 2007.

W. Cramer, A. Bondeau, F. I. Woodward, I. C. Prentice, R. A. Betts, V. Brovkin, P. M. Cox, V. Fisher, J. A. Foley, A. D. Friend, et al. Global response of terrestrial ecosystem structure and function to co2 and climate change: results from six dynamic global vegetation models. *Global change biology*, 7(4):357–373, 2001.

R. Crawford, C. Chappell, C. Thompson, and F. Rocchio. Vegetation classification of mount rainier, north cascades, and olympic national parks. plant association descriptions and identification keys. Technical report, Natural Resource Technical Report NPS/NCCN/NRTR2009/D-586. US Department of the Interior, National Park Service, Natural Resource Program Centre, Fort Collins, CO, US, 2009.

E. C. B. de Colstoun, M. H. Story, C. Thompson, K. Commisso, T. G. Smith, and J. R. Irons. National park vegetation mapping using multitemporal landsat 7 data and a decision tree classifier. *Remote Sensing of Environment*, 85(3):316–327, 2003.

R. DeFries and J. C.-W. Chan. Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. *Remote Sensing of Environment*, 74(3):503–515, 2000.

R. D. Dorn. *Vascular plants of Wyoming*, volume 2. Cheyenne, Wyoming: Mountain West Publishing 340p.-illus.. En Icones, Keys, 1992.

J. J. Feddema, K. W. Oleson, G. B. Bonan, L. O. Mearns, L. E. Buja, G. A. Meehl, and W. M. Washington. The importance of land-cover change in simulating future climates. *Science*, 310(5754):1674–1678, 2005.

M. A. Friedl and C. E. Brodley. Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, 61(3):399–409, 1997.

D. Gesch, M. Oimoen, S. Greenlee, C. Nelson, M. Steuck, and D. Tyler. The national elevation dataset. *Photogrammetric engineering and remote sensing*, 68(1):5–32, 2002.

D. Gesch, M. Oimoen, and G. Evans. Accuracy assessment of the u.s. geological survey national elevation dataset, and comparison with other large-area elevation datasetssrtm and aster. *Photogrammetric engineering and remote sensing*, 68(1): 5–32, 2014.

N. Gorelick. Google earth engine. In *AGU Fall Meeting Abstracts*, volume 1, page 04, 2012.

A. Guisan and N. E. Zimmermann. Predictive habitat distribution models in ecology. *Ecological modelling*, 135(2):147–186, 2000.

J. Hansen, M. Sato, R. Ruedy, P. Kharecha, A. Lacis, R. Miller, L. Nazarenko, K. Lo, G. Schmidt, G. Russell, et al. Climate simulations for 1880–2003 with giss modele. *Climate Dynamics*, 29(7-8):661–696, 2007.

M. Hansen, R. DeFries, J. R. Townshend, and R. Sohlberg. Global land cover classification at 1 km spatial resolution using a classification tree approach. *International journal of remote sensing*, 21(6-7):1331–1364, 2000.

K. Hayhoe, D. Cayan, C. B. Field, P. C. Frumhoff, E. P. Maurer, N. L. Miller, S. C. Moser, S. H. Schneider, K. N. Cahill, E. E. Cleland, et al. Emissions pathways, climate change, and impacts on california. *Proceedings of the National Academy of Sciences of the United States of America*, 101(34):12422–12427, 2004.

T. Hickler, K. Vohland, J. Feehan, P. A. Miller, B. Smith, L. Costa, T. Giesecke, S. Fronzek, T. R. Carter, W. Cramer, et al. Projecting the future distribution of european potential natural vegetation zones with a generalized, tree species-based dynamic vegetation model. *Global Ecology and Biogeography*, 21(1):50–63, 2012.

R. Hijmans, S. Cameron, J. Parra, P. Jones, and A. Jarvis. Worldclimglobal climate data—free climate data for ecological modeling and gis. 2015. URL `http://www.worldclim. org`.

R. J. Hijmans and C. H. Graham. The ability of climate envelope models to predict the effect of climate change on species distributions. *Global change biology*, 12(12): 2272–2281, 2006.

R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology*, 25(15):1965–1978, 2005.

D. W. Hilbert and B. Ostendorf. The utility of artificial neural networks for modelling the distribution of vegetation in past, present and future climates. *Ecological modelling*, 146(1):311–327, 2001.

C. Homer, C. Huang, L. Yang, B. Wylie, and M. Coan. Development of a 2001 national land-cover database for the united states. *Photogrammetric Engineering & Remote Sensing*, 70(7):829–840, 2004.

C. Homer, J. Dewitz, J. Fry, M. Coan, N. Hossain, C. Larson, N. Herold, A. McKerrow, J. N. VanDriel, and J. Wickham. Completion of the 2001 national land cover database for the counterminous united states. *Photogrammetric Engineering and Remote Sensing*, 73(4):337, 2007.

B. Huntley, P. M. Berry, W. Cramer, and A. P. McDonald. Special paper: modelling present and potential future ranges of some european higher plants using climate response surfaces. *Journal of Biogeography*, pages 967–1001, 1995.

B. Huntley, R. E. Green, Y. C. Collingham, J. K. Hill, S. G. Willis, P. J. Bartlein, W. Cramer, W. J. Hagemeijer, and C. J. Thomas. The performance of models relating species geographical distributions to climate is independent of trophic level. *Ecology Letters*, 7(5):417–426, 2004.

M. Hutchinson. Interpolating mean rainfall using thin plate smoothing splines. *International journal of geographical information systems*, 9(4):385–403, 1995.

R. Ihaka and R. Gentleman. R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314, 1996.

IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013. ISBN ISBN 978-1-107-66182-0. doi: 10.1017/CBO9781107415324. URL `www.climatechange2013.org`.

I. Klein, U. Gessner, and C. Kuenzer. Regional land cover mapping and change detection in central asia using modis time-series. *Applied Geography*, 35(1):219–234, 2012.

G. Krinner, N. Viovy, N. de Noblet-Ducoudré, J. Ogée, J. Polcher, P. Friedlingstein, P. Ciais, S. Sitch, and I. C. Prentice. A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, 19(1), 2005.

M. Kuhn. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

E. F. Lambin, H. J. Geist, and E. Lepers. Dynamics of land-use and land-cover change in tropical regions. *Annual review of environment and resources*, 28(1): 205–241, 2003.

T. Masui, K. Matsumoto, Y. Hijioka, T. Kinoshita, T. Nozawa, S. Ishiwatari, E. Kato, P. Shukla, Y. Yamagata, and M. Kainuma. An emission pathway for stabilization at 6 wm- 2 radiative forcing. *Climatic Change*, 109(1-2):59–76, 2011.

D. McIver and M. Friedl. Using prior probabilities in decision-tree classification of remotely sensed data. *Remote Sensing of Environment*, 81(2):253–261, 2002.

M. Meinshausen, S. J. Smith, K. Calvin, J. S. Daniel, M. Kainuma, J. Lamarque, K. Matsumoto, S. Montzka, S. Raper, K. Riahi, et al. The rcp greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic change*, 109(1-2): 213–241, 2011.

J. M. Melillo, A. D. McGuire, D. W. Kicklighter, B. Moore, C. J. Vorosmarty, and A. L. Schloss. Global climate change and terrestrial net primary production. *Nature*, 363(6426):234–240, 1993.

D. Michie, D. J. Spiegelhalter, and C. C. Taylor. *Machine learning, neural and statistical classification*. Citeseer, 1994.

T. M. Mitchell. *Machine learning. 1997*, volume 45. Burr Ridge, IL: McGraw Hill, 1997.

F. Mouillot, S. Rambal, and R. Joffre. Simulating climate change impacts on fire frequency and vegetation dynamics in a mediterranean-type ecosystem. *Global Change Biology*, 8(5):423–437, 2002.

L. Nazarenko, G. Schmidt, R. Miller, N. Tausnev, M. Kelley, R. Ruedy, G. Russell, I. Aleinov, M. Bauer, S. Bauer, et al. Future climate change under rcp emission scenarios with giss modele2. *Journal of Advances in Modeling Earth Systems*, 2015.

M. Neteler and H. Mitasova. *Open source GIS: a GRASS GIS approach*. Springer Science & Business Media, 2002.

R. G. Pearson and T. P. Dawson. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global ecology and biogeography*, 12(5):361–371, 2003.

C. Petit, T. Scudder, and E. Lambin. Quantifying processes of land-cover change by remote sensing: resettlement and rapid land-cover changes in south-eastern zambia. *International Journal of Remote Sensing*, 22(17):3435–3456, 2001.

R. A. Pielke. Land use and climate change. *Science*, 310(5754):1625–1626, 2005.

D. C. Powell, C. Johnson, E. A. Crowe, A. Wells, and D. K. Swanson. *Potential vegetation hierarchy for the Blue Mountains section of northeastern Oregon, southeastern Washington, and west-central Idaho*. US Department of Agriculture, Forest Service, Pacific Northwest Research Station, 2007.

J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

J. R. Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

J. S. Racine. Rstudio: A platform-independent ide for r and sweave. *Journal of Applied Econometrics*, 27(1):167–172, 2012.

V. Ramaswamy, O. Boucher, J. Haigh, D. Hauglustaine, J. Haywood, G. Myhre, T. Nakajima, G. Shi, S. Solomon, R. E. Betts, et al. Radiative forcing of climate change. Technical report, Pacific Northwest National Laboratory (PNNL), Richland, WA (US), 2001.

K. Riahi, S. Rao, V. Krey, C. Cho, V. Chirkov, G. Fischer, G. Kindermann, N. Nakicenovic, and P. Rafaj. Rcp 8.5a scenario of comparatively high greenhouse gas emissions. *Climatic Change*, 109(1-2):33–57, 2011.

J. Rogan, J. Franklin, D. Stow, J. Miller, C. Woodcock, and D. Roberts. Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. *Remote Sensing of Environment*, 112(5):2272–2283, 2008.

P. Romanski, L. Kotthoff, and M. L. Kotthoff. Package fselector. 2013.

R. RuleQuest. Is see5/c5. 0 better than c4. 5. *St Ive, Australia*, 2009.

O. E. Sala, F. S. Chapin, J. J. Armesto, E. Berlow, J. Bloomfield, R. Dirzo, E. Huber-Sanwald, L. F. Huenneke, R. B. Jackson, A. Kinzig, et al. Global biodiversity scenarios for the year 2100. *science*, 287(5459):1770–1774, 2000.

G. A. Schmidt, M. Kelley, L. Nazarenko, R. Ruedy, G. L. Russell, I. Aleinov, M. Bauer, S. E. Bauer, M. K. Bhat, R. Bleck, et al. Configuration and assessment of the giss modele2 contributions to the cmip5 archive. *Journal of Advances in Modeling Earth Systems*, 6(1):141–184, 2014.

S. E. Sesnie, P. E. Gessler, B. Finegan, and S. Thessler. Integrating landsat tm and srtm-dem derived variables with decision trees for habitat classification and change detection in complex neotropical environments. *Remote Sensing of Environment*, 112(5):2145–2159, 2008.

S. Sitch, C. Huntingford, N. Gedney, P. Levy, M. Lomas, S. Piao, R. Betts, P. Ciais, P. Cox, P. Friedlingstein, et al. Evaluation of the terrestrial carbon cycle, future plant geography and climate-carbon cycle feedbacks using five dynamic global vegetation models (dgvms). *Global Change Biology*, 14(9):2015–2039, 2008.

U. S. D. o. A. Soil Survey Staff, Natural Resources Conservation Service. Distribution maps of dominant soil orders. 2015. URL `http://www.nrcs.usda.gov`.

P. K. Srivastava, D. Han, M. A. Rico-Ramirez, M. Bray, and T. Islam. Selection of classification techniques for land use/land cover change investigation. *Advances in Space Research*, 50(9):1250–1265, 2012.

A. M. Thomson, K. V. Calvin, S. J. Smith, G. P. Kyle, A. Volke, P. Patel, S. Delgado-Arias, B. Bond-Lamberty, M. A. Wise, L. E. Clarke, et al. Rcp4. 5: a pathway for stabilization of radiative forcing by 2100. *Climatic Change*, 109(1-2):77–94, 2011.

D. P. Van Vuuren, E. Stehfest, M. G. den Elzen, T. Kram, J. van Vliet, S. Deetman, M. Isaac, K. K. Goldewijk, A. Hof, A. M. Beltran, et al. Rcp2. 6: exploring the possibility to keep global mean temperature increase below 2 c. *Climatic Change*, 109(1-2):95–116, 2011.

A. Weiss. Topographic position and landforms analysis. In *Poster presentation, ESRI User Conference, San Diego, CA*, pages 200–200, 2001.

M. S. Wigmosta, L. W. Vail, and D. P. Lettenmaier. A distributed hydrology-vegetation model for complex terrain. *Water resources research*, 30(6):1665–1679, 1994.

I. Witten, S. Cunningham, G. Holmes, R. McQueen, and L. Smith. Practical machine learning and its potential application to problems in agriculture. In *Proc New Zealand Computer Conference*, volume 1, pages 308–325, Auckland, New Zealand, 1993.

I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

L. Yang, C. Homer, K. Hegge, C. Huang, B. Wylie, and B. Reed. A landsat 7 scene selection strategy for a national land cover database. In *Geoscience and Remote Sensing Symposium, 2001. IGARSS'01. IEEE 2001 International*, volume 3, pages 1123–1125. IEEE, 2001.

L. Yang, G. Xian, J. M. Klaver, and B. Deal. Urban land-cover change detection through sub-pixel imperviousness mapping using remotely sensed data. *Photogrammetric Engineering & Remote Sensing*, 69(9):1003–1010, 2003.